ANALOG INTEGRATED CIRCUIT DESIGN

SECOND EDITION

Tony Chan Carusone | David A. Johns | Kenneth W. Martin

ANALOG INTEGRATED CIRCUIT DESIGN

Tony Chan Carusone
David A. Johns
Kenneth W. Martin

WILEY

John Wiley & Sons, Inc.

VP and Publisher	Don Fowley
Associate Publisher	Dan Sayre
Editorial Assistant	Charlotte Cerf
Senior Marketing Manager	Christopher Ruel
Senior Production Manager	Janis Soo
Senior Production Editor	Joyce Poh

This book was set in 9.5/11.5 Times New Roman PSMT by MPS Limited, a Macmillan Company, and printed and bound by RRD Von Hoffman. The cover was printed by RRD Von Hoffman.

This book is printed on acid free paper.

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

Copyright © 2012 John Wiley & Sons, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc. 222 Rosewood Drive, Danvers, MA 01923, website www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, (201)748-6011, fax (201)748-6008, website http://www.wiley.com/go/permissions.

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at www.wiley.com/go/returnlabel. Outside of the United States, please contact your local representative.

Library of Congress Cataloging-in-Publication Data
Carusone, Tony Chan.
Analog integrated circuit design / Tony Chan Carusone, David A. Johns, Kenneth W. Martin. —2nd ed.
p. cm.
Includes index.
Prev ed. listed under David A. Johns.
ISBN 978-0-470-77010-8 (pbk.)
I. Johns, David, 1958 II. Martin, Kenneth W. (Kenneth William) 1952 III. Johns, David, 1958- Analog integrated circuit design. IV. Title.
TK7874.J65 2011
621.3815—dc23

2011039275

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

To Soo, Brayden, Teague, and Senna
To Cecilia, Christopher, Timothy, and Victoria
To Elisabeth and Jeremy

Preface

It has long been predicted that there would soon be little need for analog circuitry because the world increasingly relies on digital signals, yet the need for good analog circuit design remains strong. Many applications have indeed replaced analog circuitry with their digital counterparts (such as digital audio). However, when digitizing physical signals, analog-to-digital and digital-to-analog converters are always needed, together with their associated anti-aliasing and reconstruction filters. In addition, new applications continue to appear; their requirements demand the use of high-performance analog front ends, such as digital communication over wireline and wireless channels and microsensor interfaces. Also, as integrated circuits integrate more functionality, it is much more likely that at least some portion of a modern integrated circuit will include analog circuitry to interface to the real world. Moreover, the continued scaling of digital circuits has led to the emergence of new problems that require analog solutions, such as on-chip power management and the generation of stable clock signals. Although it may constitute only a small portion of total chip area, analog circuitry is often the limiting factor on overall system performance and the most difficult part of the IC to design. As a result, a strong industrial need for analog circuit designers continues. The purpose of this book is to help develop excellent analog circuit designers by presenting a concise treatment of the wide array of knowledge required by an integrated circuit designer.

This book strives to quash the notion that the design and test of high-performance analog circuits are "mystical arts." Whereas digital design is relatively systematic, analog design appears to be much more based upon intuition and experience. Analog testing may sometimes seem to depend more upon the time of day and phase of the moon than on concrete electrical properties. But these thoughts about analog circuits usually occur when one is not familiar with the many fundamentals required to create high-performance analog circuits. This book helps to take the mystery out of analog integrated circuit design. Although many circuits and techniques are described, the most important design principles are emphasized throughout this book. Physical and intuitive explanations are given, and although mathematical quantitative analyses of many circuits have necessarily been presented, one must not miss seeing the forest for the trees. In other words, this book attempts to present the critical underlying concepts without becoming entangled in tedious and overcomplicated circuit analyses.

NEW TO THIS EDITION

This, the second edition of Analog Integrated Circuit Design, has new material to make it more accessible to beginners in the field while retaining the depth, detail, and intuitive approach that made the first edition a favorite reference among experienced designers. Two new chapters have been added early in the text: Chapter 4, dedicated to the frequency response of analog integrated circuits, provides a review of frequency-domain analysis and single-stage amplifier response; Chapter 5 covers the basic theory of feedback amplifiers. The conventional categorization and dissection of feedback amplifiers according to their topology is by and large forgone in favor of an intuitive, practical, yet analytical approach that is based on the practices of experienced analog designers. These new chapters make the second edition well-suited to the teaching of analog integrated circuit design at both the undergraduate and graduate levels, while still allowing it to serve as a comprehensive reference for practicing engineers.

The first edition of Analog Integrated Circuit Design was written roughly 15 years before the second, and the field changed considerably in the intervening years necessitating significant updates to reflect advances in

technology and engineering practice. For example, material on CMOS integrated circuit device modeling, processing, and layout in Chapters 1 and 2 has been updated and expanded to cover effects that are of tremendous importance to analog designers using modern fabrication technologies. New and expanded topics include modeling MOS subthreshold operation and mobility degradation in Chapter 1, and proximity effects and mismatch both covered under the subheading "Variability" in Chapter 2. Also in Chapter 1, the increasingly important role of simulation in the early phases of analog design is reflected by relating MOS parameters to the results of practical simulations. Simulation examples have been added throughout the text, particularly in the early chapters. Circuits and architectures whose fundamental importance have emerged over the past decade have been added such as voltage regulators (in Chapter 7) and the 1.5-bit-per-stage pipelined A/D converter (in Chapter 17). New circuit topologies specifically suited to low-voltage operation are presented, such as a low-voltage bandgap reference circuit in Chapter 7. Nonlinearity and dynamic range are now presented in Chapter 9 alongside noise, highlighting their fundamental interrelationship. New study problems have been added throughout the text and numerical examples have been updated to reflect the realities of modern fabrication technologies.

This edition has also been updated to accommodate today's varying pedagogical approaches toward the teaching of bipolar devices and circuits. Material on bipolar devices and circuits, which was scattered over several chapters of the first edition, has been combined into Chapter 8 of this edition. The reorganization permits undergraduate-level instructors and readers to either incorporate or omit the material at their discretion. In the later chapters, readers are assumed to have experience with analog design, hence bipolar and BiCMOS circuits are presented alongside CMOS circuits, as in the first edition.

Finally, Chapter 19 on phase-locked loops (PLLs) has been rewritten. When the first edition was released, it was one of the first analog circuit texts to elucidate the design of integrated circuit PLLs. Today, fully-integrated PLLs have become a basic building block of both analog and mostly-digital integrated circuits. As such, the material has become standard fare at the graduate level, and increasingly at the undergraduate level too. Chapter 19 now provides a thorough treatment of jitter and phase noise, major performance metrics in the design of modern PLLs and clocked systems.

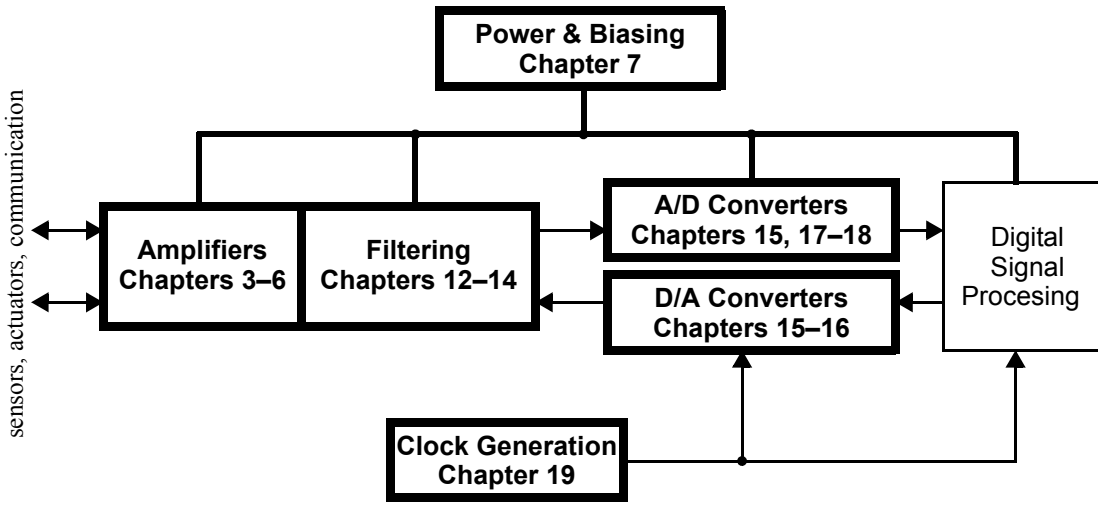
INTENDED AUDIENCE

This book is intended for use as a senior-undergraduate and graduate-level textbook, and as a reference for practicing engineers. To appreciate the material in this book, it is expected that the reader has had at least one basic introductory course in electronics. Specifically, the reader should be familiar with the concept of small-signal analysis and have been exposed to basic transistor circuits. In addition, the reader should be have been exposed to Fourier and Laplace transforms. Some prior knowledge of discrete-time signal processing is important for the later chapters. Although all of these topics are reviewed, background in these areas will benefit the reader significantly.

The chapters of this book have intentionally been made mostly independent so that some chapters can be covered while others are skipped. Also, it has been found to be very easy to change the order of presentation. For example, if readers have a good modelling background they might skip Chapter 1, and if their discrete-time knowledge is good Chapter 13 might he assigned only as review. We believe that such flexibility is essential in presenting textbooks for the later years of study.

The material in this book can be used for a few courses. A second undergraduate course in electronics typically has frequency response and feedback, as its major topics. For such a course, Chapters 1, 3, 4 and 5 may be assigned. Some advanced modeling from Chapter 1 may be omitted and replaced with selected topics from Chapters 2 and 6 at the instructor's discretion. A senior-level undergraduate course in analog integrated circuits assigns Chapters 1, 2, 6, and 7, with Chapters 3–5 serving as a useful reference for those students requiring extra review. Chapter 8 may be included in any course that covers bipolar processing and devices.

A senior undergraduate or entry-level graduate course on analog signal processing may use Chapters 9–14. A graduate-level course on data converters will focus upon Chapters 15–18, drawing upon the earlier chapters as



needed for supplementary material. Finally, Chapter 19 may be used for a graduate level course on phase locked loops. Naturally there is considerable variability in the specific readings assigned by different instructors, particularly at the graduate level. This variability is recognized in the basic organization of the book.

A secondary audience for this book includes recently graduated electrical engineers who wish to rapidly increase their knowledge of modern analog circuit design techniques. In fact, much of the material covered in this text was originally taught and refined over many years in popular short courses offered to working engineers who realized the importance of upgrading their knowledge in analog circuit design. For this audience, we have put effort into highlighting the most important considerations when designing the various circuits. We have also tried to include modern, well-designed examples and references to primary sources for further study.

TEXT OUTLINE

Analog integrated circuits are critical blocks that permeate complex electronic systems. Analog circuits inevitably arise whenever those systems must interact with the analog world of sensors or actuators (including antennas, cameras, microphones, speakers, displays, lighting, motors, and many others), and when they must communicate using anything but the most rudimentary digital signals. A typical system is illustrated in the figure. The blocks covered in some detail in this text are highlighted, and the corresponding chapters referenced. Chapters describing the design of amplifiers, and all chapters not explicitly referenced in the figure, are foundational and relevant to the implementation of many analog and mixed-signal systems. The table of contents provides a catalog of the book's main topics. What follows here is a very brief summary of each chapter.

In Chapter 1, the basic physical behavior and modelling of diodes, MOS transistors, and integrated circuit capacitors and resistors are covered. Here, many of the modelling equations are derived to give the reader some appreciation of model parameters and how they are affected by processes parameters. Diode and MOSFET models are summarized in a table format for quick reference.

In Chapter 2, issues associated with the manufacturing of an integrated circuit are discussed. Emphasis is placed on CMOS fabrication. In addition to the provided background, issues that are of particular importance to analog designers are emphasized, such as variability (including random mismatch) layout rules and best practices.

Fundamental building blocks of analog integrated circuits are discussed in Chapter–3, specifically, MOS current mirrors and single-stage amplifiers, concluding with the basic MOS differential pair. A point to note here is that only active-load amplifiers are considered since these are prevalent in integrated circuits.

Chapter 4 provides an introductory view of the frequency response of electronic circuits. It begins with fundamental material on frequency response, establishing definitions and notation for the following chapters. Then, the frequency response of elementary CMOS analog building blocks is presented. Along the way, fundamental topics are presented including the Miller effect and the method of zero-value time-constants.

Feedback amplifiers are introduced in Chapter 5. Loop gain and phase margin are defined. Basic concepts are illustrated using generic analyses of first- and second-order feedback systems. At the end of the chapter, the analysis is applied to common CMOS feedback circuits.

In Chapter 6, the fundamental principles of basic opamp design are presented. To illustrate many of these principles, the design of a classic two-stage CMOS opamp is first thoroughly discussed. Proper biasing and device sizing strategies are covered. Compensation is introduced and a systematic procedure for compensation is described. Then, advanced current-mirror approaches are discussed, followed by two opamps that make use of them: the folded-cascode and current mirror opamps. Finally, fully differential opamps are presented, as they are used in many modern industrial applications where high speed and low noise are important considerations.

Biasing, reference, and regulators are presented in Chapter 7. Any reader that wishes to design a real and complete opamp circuit should be aware of the attendant issues covered here. The later sections on bandgap references and voltage regulators may not be essential to all readers.

Chapter 8 provides a comprehensive summary of bipolar devices and circuits. It includes the basics of device modeling, fabrication, and fundamental circuit blocks such as current mirrors and gain stages. The reader may wish to read sections of this chapter alongside the corresponding material for MOS transistors presented in Chapters 1–7.

Noise analysis and modelling and linearity are discussed in Chapter 9. Here, we assume the reader has not previously been exposed to random-signal analysis, and thus basic concepts in analyzing random signals are first presented. Noise models are then presented for basic circuit elements. A variety of circuits are analyzed from a noise perspective giving the reader some experience in noise analysis. Finally, the concept of dynamic range is introduced as a fundamental specification of most any analog circuit, and the basic measures of linearity are defined.

In Chapter 7, comparator design is discussed. Comparators are perhaps the second most common analog building block after opamps. Here, the practical limitations of comparators are described as well as circuit techniques to improve performance. In addition, examples of modern high-speed comparators are presented.

In Chapter 11, some additional analog building blocks are covered. Specifically, sample-and-hold circuits and translinear gain and multiplier circuits are presented. By the end of this chapter, all the main analog building blocks have been covered (with the possible exception of voltage-controlled oscillators) and the remaining material in the text deals with more system-level analog considerations.

Continuous-time filters are the focus of Chapter 12. After a brief introduction to first- and second-order filters, transconductance-C filters are described. CMOS, bipolar, and BiCMOS approaches are covered. Active-RC filters are then presented, followed by some tuning approaches. Finally, a brief introduction to complex analog signal processing and complex filters is included.

The basics of discrete-time signals and filters are presented in Chapter 13. This material is essential for understanding the operation of many analog circuits such as switched-capacitor filters and oversampling converters. The approach taken here is to show the close relationship between the Z-transform and the Laplace transform, thereby building on the reader's experience in the continuous-time domain.

In Chapter 14, the basics of switched-capacitor circuits are described. Switched-capacitor techniques are a common approach for realizing integrated filters due to their high degree of accuracy and linearity. The chapter concludes with a description of other switched-capacitor circuits. such as gain stages, modulators, and voltage-controlled oscillators.

In Chapter 15, the fundamentals of data converters are presented. Ideal converters and the properties of quantization noise are discussed first. Signed codes are then presented, and the chapter concludes with a discussion of performance limitations and metrics.

Popular Nyquist-rate D/A architectures are discussed in Chapter 16 and various approaches for realizing Nyquist-rate A/D converters are described in Chapter 17. The importance of data converters cannot be overemphasized in today's largely digital world, and these two chapters discuss the main advantages and design issues of many modern approaches.

Oversampling conveners are presented separately in Chapter 18 due to the large amount of signalprocessing concepts needed to properly describe these converters. Here, digital issues (such as decimation filters) are also presented since good overall system knowledge is needed to properly design these types of converters. In addition, practical issues and advanced approaches (such as the use of bandpass and multibit converters) are also discussed. This chapter concludes with a third-order A/D converter example.

Finally, the text concludes with phase-locked loops (PLLs) in Chapter 19. The chapter first provides a bigpicture overview of PLLs. A more rigorous treatment follows, including small-signal analysis and noise analysis in both the time domain (jitter) and frequency domain (phase noise). Performance metrics and design procedures are included.

USING THE BOOK AND WEBSITE

SPICE simulation examples are an important feature of the book. Passages annotated with the boxed icon shown here indicate that a SPICE simulation may be performed either as an essential part of the problem, or to corroborate the results of a hand analysis. Many of the problems and examples in this book rely upon the fictitious CMOS process technologies whose parameters are summarized in Table 1.5. SPICE model files corresponding to each of these fictitious technologies are provided on the companion website, **www.analogicdesign.com**. Also there are many netlists that may be used for the simulations. The results they provide should roughly corroborate hand analyses performed using the parameters in Table 1.5. However, simulation results *never* provide precise agreement. In fact, simulated results may differ from the results of a hand analysis by as much as 50%! This is a reality of analog design, and the SPICE examples in this book are no exception. This is, of itself, a valuable lesson to the student of analog design. It illustrates, through practice, those tasks to which hand analysis and simulation are best suited.

End-of-chapter problems are organized by the subsection to which they pertain. For example, if one wishes to practice only those problems pertaining to current mirror opamps, one may proceed directly to Section 6.11.5.

Key points throughout the text are emphasized using highlighted boxes in the margins, as shown here. These key points are collected and listed at the end of each chapter.

Key Point: *Key points throughout the text are emphasized using separate highlighted boxes in the margins. These key points are collected and listed at the end of each chapter as a study aid.*

ACKNOWLEDGEMENTS

The authors would like to acknowledge the many colleagues who participated in short courses, during which much of the material for this text was originally taught and refined. In particular, Gabor C. Temes is acknowledged as well as instructors Jim McCreary and Bill Black. Many students and anonymous reviewers also diligently reviewed and provided corrections to the manuscript—their help is gratefully acknowledged. In addition, the authors acknowledge that much of the material and many of the concepts originated from work with practicing engineers over the years, with as well as in the publications cited in the references section at the end of each chapter. As much as possible, appropriate references for original concepts are cited in the text, but the authors have been working in the area of analog circuits for so many years that often the original sources of popular and important concepts have been forgotten. For any reference omissions, they sincerely apologize.

Contents

CHAPTER 1 INTEGRATED-CIRCUIT DEVICES AND MODELLING 1

1.1 Semiconductors and pn Junctions 1
- 1.1.1 Diodes 2
- 1.1.2 Reverse-Biased Diodes 4
- 1.1.3 Graded Junctions 7
- 1.1.4 Large-Signal Junction Capacitance 9
- 1.1.5 Forward-Biased Junctions 10
- 1.1.6 Junction Capacitance of Forward-Biased Diode 11
- 1.1.7 Small-Signal Model of a Forward-Biased Diode 12
- 1.1.8 Schottky Diodes 13

1.2 MOS Transistors 14
- 1.2.1 Symbols for MOS Transistors 15
- 1.2.2 Basic Operation 16
- 1.2.3 Large-Signal Modelling 21
- 1.2.4 Body Effect 24
- 1.2.5 p-Channel Transistors 24
- 1.2.6 Low-Frequency Small-Signal Modelling in the Active Region 25
- 1.2.7 High-Frequency Small-Signal Modelling in the Active Region 30
- 1.2.8 Small-Signal Modelling in the Triode and Cutoff Regions 33
- 1.2.9 Analog Figures of Merit and Trade-offs 36

1.3 Device Model Summary 38
- 1.3.1 Constants 38
- 1.3.2 Diode Equations 39
- 1.3.3 MOS Transistor Equations 40

1.4 Advanced MOS Modelling 42
- 1.4.1 Subthreshold Operation 42
- 1.4.2 Mobility Degradation 44
- 1.4.3 Summary of Subthreshold and Mobility Degradation Equations 47
- 1.4.4 Parasitic Resistances 47
- 1.4.5 Short-Channel Effects 48
- 1.4.6 Leakage Currents 49

1.5 SPICE Modelling Parameters 50
- 1.5.1 Diode Model 50
- 1.5.2 MOS Transistors 51
- 1.5.3 Advanced SPICE Models of MOS Transistors 51

1.6 Passive Devices 54
- 1.6.1 Resistors 54
- 1.6.2 Capacitors 58

1.7 Appendix 60
1.7.1 Diode Exponential Relationship 60
1.7.2 Diode-Diffusion Capacitance 62
1.7.3 MOS Threshold Voltage and the Body Effect 64
1.7.4 MOS Triode Relationship 66
1.8 Key Points 68
1.9 References 69
1.10 Problems 69

CHAPTER 2 PROCESSING AND LAYOUT 73

2.1 CMOS Processing 73
2.1.1 The Silicon Wafer 73
2.1.2 Photolithography and Well Definition 74
2.1.3 Diffusion and Ion Implantation 76
2.1.4 Chemical Vapor Deposition and Defining the Active Regions 78
2.1.5 Transistor Isolation 78
2.1.6 Gate-Oxide and Threshold-Voltage Adjustments 81
2.1.7 Polysilicon Gate Formation 82
2.1.8 Implanting the Junctions, Depositing SiO_2, and Opening Contact Holes 82
2.1.9 Annealing, Depositing and Patterning Metal, and Overglass Deposition 84
2.1.10 Additional Processing Steps 84
2.2 CMOS Layout and Design Rules 86
2.2.1 Spacing Rules 86
2.2.2 Planarity and Fill Requirements 94
2.2.3 Antenna Rules 94
2.2.4 Latch-Up 95
2.3 Variability and Mismatch 96
2.3.1 Systematic Variations Including Proximity Effects 96
2.3.2 Process Variations 98
2.3.3 Random Variations and Mismatch 99
2.4 Analog Layout Considerations 103
2.4.1 Transistor Layouts 103
2.4.2 Capacitor Matching 104
2.4.3 Resistor Layout 107
2.4.4 Noise Considerations 109
2.5 Key Points 113
2.6 References 114
2.7 Problems 114

CHAPTER 3 BASIC CURRENT MIRRORS AND SINGLE-STAGE AMPLIFIERS 117

3.1 Simple CMOS Current Mirror 118
3.2 Common-Source Amplifier 120
3.3 Source-Follower or Common-Drain Amplifier 122

3.4 Common-Gate Amplifier 124
3.5 Source-Degenerated Current Mirrors 127
3.6 Cascode Current Mirrors 129
3.7 Cascode Gain Stage 131
3.8 MOS Differential Pair and Gain Stage 135
3.9 Key Points 138
3.10 References 139
3.11 Problems 139

CHAPTER 4 FREQUENCY RESPONSE OF ELECTRONIC CIRCUITS 144

4.1 Frequency Response of Linear Systems 144
4.1.1 Magnitude and Phase Response 145
4.1.2 First-Order Circuits 147
4.1.3 Second-Order Low-Pass Transfer Functions with Real Poles 154
4.1.4 Bode Plots 157
4.1.5 Second-Order Low-Pass Transfer Functions with Complex Poles 163
4.2 Frequency Response of Elementary Transistor Circuits 165
4.2.1 High-Frequency MOS Small-Signal Model 165
4.2.2 Common-Source Amplifier 166
4.2.3 Miller Theorem and Miller Effect 169
4.2.4 Zero-Value Time-Constant Analysis 173
4.2.5 Common-Source Design Examples 176
4.2.6 Common-Gate Amplifier 179
4.3 Cascode Gain Stage 181
4.4 Source-Follower Amplifier 187
4.5 Differential Pair 193
4.5.1 High-Frequency T-Model 193
4.5.2 Symmetric Differential Amplifier 194
4.5.3 Single-Ended Differential Amplifier 195
4.5.4 Differential Pair with Active Load 196
4.6 Key Points 197
4.7 References 198
4.8 Problems 199

CHAPTER 5 FEEDBACK AMPLIFIERS 204

5.1 Ideal Model of Negative Feedback 204
5.1.1 Basic Definitions 204
5.1.2 Gain Sensitivity 205
5.1.3 Bandwidth 207
5.1.4 Linearity 207
5.1.5 Summary 208
5.2 Dynamic Response of Feedback Amplifiers 208
5.2.1 Stability Criteria 209
5.2.2 Phase Margin 211

5.3 First- and Second-Order Feedback Systems 213
5.3.1 First-Order Feedback Systems 213
5.3.2 Second-Order Feedback Systems 217
5.3.3 Higher-Order Feedback Systems 220
5.4 Common Feedback Amplifiers 220
5.4.1 Obtaining the Loop Gain, L(s) 222
5.4.2 Non-Inverting Amplifier 226
5.4.3 Transimpedance (Inverting) Amplifiers 231
5.5 Summary of Key Points 235
5.6 References 235
5.7 Problems 236

CHAPTER 6 BASIC OPAMP DESIGN AND COMPENSATION 242

6.1 Two-Stage CMOS Opamp 242
6.1.1 Opamp Gain 243
6.1.2 Frequency Response 245
6.1.3 Slew Rate 249
6.1.4 *n*-Channel or *p*-Channel Input Stage 252
6.1.5 Systematic Offset Voltage 252
6.2 Opamp Compensation 254
6.2.1 Dominant-Pole Compensation and Lead Compensation 254
6.2.2 Compensating the Two-Stage Opamp 255
6.2.3 Making Compensation Independent of Process and Temperature 259
6.3 Advanced Current Mirrors 261
6.3.1 Wide-Swing Current Mirrors 261
6.3.2 Enhanced Output-Impedance Current Mirrors and Gain Boosting 263
6.3.3 Wide-Swing Current Mirror with Enhanced Output Impedance 266
6.3.4 Current-Mirror Symbol 267
6.4 Folded-Cascode Opamp 268
6.4.1 Small-Signal Analysis 270
6.4.2 Slew Rate 272
6.5 Current Mirror Opamp 275
6.6 Linear Settling Time Revisited 279
6.7 Fully Differential Opamps 281
6.7.1 Fully Differential Folded-Cascode Opamp 283
6.7.2 Alternative Fully Differential Opamps 284
6.7.3 Low Supply Voltage Opamps 286
6.8 Common-Mode Feedback Circuits 288
6.9 Summary of Key Points 292
6.10 References 293
6.11 Problems 294

CHAPTER 7 BIASING, REFERENCES, AND REGULATORS 302

7.1 Analog Integrated Circuit Biasing 302
7.1.1 Bias Circuits 303

7.1.2 Reference Circuits 305
7.1.3 Regulator Circuits 306
7.2 Establishing Constant Transconductance 307
7.2.1 Basic Constant-Transconductance Circuit 307
7.2.2 Improved Constant-Transconductance Circuits 309
7.3 Establishing Constant Voltages and Currents 310
7.3.1 Bandgap Voltage Reference Basics 310
7.3.2 Circuits for Bandgap References 314
7.3.3 Low-Voltage Bandgap Reference 319
7.3.4 Current Reference 320
7.4 Voltage Regulation 321
7.4.1 Regulator Specifications 322
7.4.2 Feedback Analysis 322
7.4.3 Low Dropout Regulators 324
7.5 Summary of Key Points 327
7.6 References 327
7.7 Problems 328

CHAPTER 8 BIPOLAR DEVICES AND CIRCUITS 331

8.1 Bipolar-Junction Transistors 331
8.1.1 Basic Operation 331
8.1.2 Analog Figures of Merit 341
8.2 Bipolar Device Model Summary 344
8.3 SPICE Modeling 345
8.4 Bipolar and BICMOS Processing 346
8.4.1 Bipolar Processing 346
8.4.2 Modern SiGe BiCMOS HBT Processing 347
8.4.3 Mismatch in Bipolar Devices 348
8.5 Bipolar Current Mirrors and Gain Stages 349
8.5.1 Current Mirrors 349
8.5.2 Emitter Follower 350
8.5.3 Bipolar Differential Pair 353
8.6 Appendix 356
8.6.1 Bipolar Transistor Exponential Relationship 356
8.6.2 Base Charge Storage of an Active BJT 359
8.7 Summary of Key Points 359
8.8 References 360
8.9 Problems 360

CHAPTER 9 NOISE AND LINEARITY ANALYSIS AND MODELLING 363

9.1 Time-Domain Analysis 363
9.1.1 Root Mean Square (rms) Value 364
9.1.2 SNR 365
9.1.3 Units of dBm 365
9.1.4 Noise Summation 366

9.2 Frequency-Domain Analysis 367
9.2.1 Noise Spectral Density 367
9.2.2 White Noise 369
9.2.3 1/f, or Flicker, Noise 370
9.2.4 Filtered Noise 371
9.2.5 Noise Bandwidth 373
9.2.6 Piecewise Integration of Noise 375
9.2.7 1/f Noise Tangent Principle 377
9.3 Noise Models for Circuit Elements 377
9.3.1 Resistors 378
9.3.2 Diodes 378
9.3.3 Bipolar Transistors 380
9.3.4 MOSFETS 380
9.3.5 Opamps 382
9.3.6 Capacitors and Inductors 382
9.3.7 Sampled Signal Noise 384
9.3.8 Input-Referred Noise 384
9.4 Noise Analysis Examples 387
9.4.1 Opamp Example 387
9.4.2 Bipolar Common-Emitter Example 390
9.4.3 CMOS Differential Pair Example 392
9.4.4 Fiber-Optic Transimpedance Amplifier Example 395
9.5 Dynamic Range Performance 397
9.5.1 Total Harmonic Distortion (THD) 398
9.5.2 Third-Order Intercept Point (IP3) 400
9.5.3 Spurious-Free Dynamic Range (SFDR) 402
9.5.4 Signal-to-Noise and Distortion Ratio (SNDR) 404
9.6 Key Points 405
9.7 References 406
9.8 Problems 406

CHAPTER 10 COMPARATORS 413

10.1 Comparator Specifications 413
10.1.1 Input Offset and Noise 413
10.1.2 Hysteresis 414
10.2 Using an Opamp for a Comparator 415
10.2.1 Input-Offset Voltage Errors 417
10.3 Charge-Injection Errors 418
10.3.1 Making Charge-Injection Signal Independent 421
10.3.2 Minimizing Errors Due to Charge-Injection 421
10.3.3 Speed of Multi-Stage Comparators 424
10.4 Latched Comparators 426
10.4.1 Latch-Mode Time Constant 427
10.4.2 Latch Offset 430

10.5 Examples of CMOS and BiCMOS Comparators 431
10.5.1 Input-Transistor Charge Trapping 435
10.6 Examples of Bipolar Comparators 437
10.7 Key Points 439
10.8 References 440
10.9 Problems 440

CHAPTER 11 SAMPLE-AND-HOLD AND TRANSLINEAR CIRCUITS 444

11.1 Performance of Sample-and-Hold Circuits 444
11.1.1 Testing Sample and Holds 445
11.2 MOS Sample-and-Hold Basics 446
11.3 Examples of CMOS S/H Circuits 452
11.4 Bipolar and BiCMOS Sample-and-Holds 456
11.5 Translinear Gain Cell 460
11.6 Translinear Multiplier 462
11.7 Key Points 464
11.8 References 465
11.9 Problems 466

CHAPTER 12 CONTINUOUS-TIME FILTERS 469

12.1 Introduction to Continuous-Time Filters 469
12.1.1 First-Order Filters 470
12.1.2 Second-Order Filters 470
12.2 Introduction to G_m-C Filters 471
12.2.1 Integrators and Summers 472
12.2.2 Fully Differential Integrators 474
12.2.3 First-Order Filter 475
12.2.4 Biquad Filter 477
12.3 Transconductors Using Fixed Resistors 479
12.4 CMOS Transconductors Using Triode Transistors 484
12.4.1 Transconductors Using a Fixed-Bias Triode Transistor 484
12.4.2 Transconductors Using Varying Bias-Triode Transistors 486
12.4.3 Transconductors Using Constant Drain-Source Voltages 491
12.5 CMOS Transconductors Using Active Transistors 493
12.5.1 CMOS Pair 493
12.5.2 Constant Sum of Gate-Source Voltages 494
12.5.3 Source-Connected Differential Pair 495
12.5.4 Inverter-Based 495
12.5.5 Differential-Pair with Floating Voltage Sources 497
12.5.6 Bias-Offset Cross-Coupled Differential Pairs 499
12.6 Bipolar Transconductors 500
12.6.1 Gain-Cell Transconductors 500
12.6.2 Transconductors Using Multiple Differential Pairs 501

12.7 BiCMOS Transconductors 506
12.7.1 Tunable MOS in Triode 506
12.7.2 Fixed-Resistor Transconductor with a Translinear Multiplier 507
12.7.3 Fixed Active MOS Transconductor with a Translinear Multiplier 508
12.8 Active RC and MOSFET-C Filters 509
12.8.1 Active RC Filters 510
12.8.2 MOSFET-C Two-Transistor Integrators 512
12.8.3 Four-Transistor Integrators 515
12.8.4 R-MOSFET-C Filters 521
12.9 Tuning Circuitry 516
12.9.1 Tuning Overview 517
12.9.2 Constant Transconductance 519
12.9.3 Frequency Tuning 520
12.9.4 Q-Factor Tuning 522
12.9.5 Tuning Methods Based on Adaptive Filtering 523
12.10 Introduction to Complex Filters 525
12.10.1 Complex Signal Processing 525
12.10.2 Complex Operations 526
12.10.3 Complex Filters 527
12.10.4 Frequency-Translated Analog Filters 528
12.11 Key Points 531
12.12 References 532
12.13 Problems 534

CHAPTER 13 DISCRETE-TIME SIGNALS 537

13.1 Overview of Some Signal Spectra 537
13.2 Laplace Transforms of Discrete-Time Signals 537
13.2.1 Spectra of Discrete-Time Signals 540
13.3 z-Transform 541
13.4 Downsampling and Upsampling 543
13.5 Discrete-Time Filters 545
13.5.1 Frequency Response of Discrete-Time Filters 545
13.5.2 Stability of Discrete-Time Filters 548
13.5.3 IIR and FIR Filters 550
13.5.4 Bilinear Transform 550
13.6 Sample-and-Hold Response 552
13.7 Key Points 554
13.8 References 555
13.9 Problems 555

CHAPTER 14 SWITCHED-CAPACITOR CIRCUITS 557

14.1 Basic Building Blocks 557
14.1.1 Opamps 557

14.1.2 Capacitors 558
14.1.3 Switches 558
14.1.4 Nonoverlapping Clocks 559
14.2 Basic Operation and Analysis 560
14.2.1 Resistor Equivalence of a Switched Capacitor 560
14.2.2 Parasitic-Sensitive Integrator 560
14.2.3 Parasitic-Insensitive Integrators 565
14.2.4 Signal-Flow-Graph Analysis 569
14.3 Noise in Switched-Capacitor Circuits 570
14.4 First-Order Filters 572
14.4.1 Switch Sharing 575
14.4.2 Fully Differential Filters 575
14.5 Biquad Filters 577
14.5.1 Low-Q Biquad Filter 577
14.5.2 High-Q Biquad Filter 581
14.6 Charge Injection 585
14.7 Switched-Capacitor Gain Circuits 588
14.7.1 Parallel Resistor-Capacitor Circuit 588
14.7.2 Resettable Gain Circuit 588
14.7.3 Capacitive-Reset Gain Circuit 591
14.8 Correlated Double-Sampling Techniques 593
14.9 Other Switched-Capacitor Circuits 594
14.9.1 Amplitude Modulator 594
14.9.2 Full-Wave Rectifier 595
14.9.3 Peak Detectors 596
14.9.4 Voltage-Controlled Oscillator 596
14.9.5 Sinusoidal Oscillator 598
14.10 Key Points 600
14.11 References 601
14.12 Problems 602

CHAPTER 15 DATA CONVERTER FUNDAMENTALS 606

15.1 Ideal D/A Converter 606
15.2 Ideal A/D Converter 608
15.3 Quantization Noise 609
15.3.1 Deterministic Approach 609
15.3.2 Stochastic Approach 610
15.4 Signed Codes 612
15.5 Performance Limitations 614
15.5.1 Resolution 614
15.5.2 Offset and Gain Error 615
15.5.3 Accuracy and Linearity 615
15.6 Key Points 620
15.7 References 620
15.8 Problems 620

CHAPTER 16 NYQUIST-RATE D/A CONVERTERS 623

16.1 Decoder-Based Converters 623
16.1.1 Resistor String Converters 623
16.1.2 Folded Resistor-String Converters 625
16.1.3 Multiple Resistor-String Converters 625
16.1.4 Signed Outputs 627
16.2 Binary-Scaled Converters 628
16.2.1 Binary-Weighted Resistor Converters 629
16.2.2 Reduced-Resistance-Ratio Ladders 630
16.2.3 R-2R-Based Converters 630
16.2.4 Charge-Redistribution Switched-Capacitor Converters 632
16.2.5 Current-Mode Converters 633
16.2.6 Glitches 633
16.3 Thermometer-Code Converters 634
16.3.1 Thermometer-Code Current-Mode D/A Converters 636
16.3.2 Single-Supply Positive-Output Converters 637
16.3.3 Dynamically Matched Current Sources 638
16.4 Hybrid Converters 640
16.4.1 Resistor-Capacitor Hybrid Converters 640
16.4.2 Segmented Converters 640
16.5 Key Points 642
16.6 References 643
16.7 Problems 643

CHAPTER 17 NYQUIST-RATE A/D CONVERTERS 646

17.1 Integrating Converters 646
17.2 Successive-Approximation Converters 650
17.2.1 DAC-Based Successive Approximation 652
17.2.2 Charge-Redistribution A/D 653
17.2.3 Resistor-Capacitor Hybrid 658
17.2.4 Speed Estimate for Charge-Redistribution Converters 658
17.2.5 Error Correction in Successive-Approximation Converters 659
17.2.6 Multi-Bit Successive-Approximation 662
17.3 Algorithmic (or Cyclic) A/D Converter 662
17.3.1 Ratio-Independent Algorithmic Converter 662
17.4 Pipelined A/D Converters 665
17.4.1 One-Bit-Per-Stage Pipelined Converter 667
17.4.2 1.5 Bit Per Stage Pipelined Converter 669
17.4.3 Pipelined Converter Circuits 672
17.4.4 Generalized k-Bit-Per-Stage Pipelined Converters 673
17.5 Flash Converters 673
17.5.1 Issues in Designing Flash A/D Converters 675
17.6 Two-Step A/D Converters 677
17.6.1 Two-Step Converter with Digital Error Correction 679
17.7 Interpolating A/D Converters 680

17.8 Folding A/D Converters 683
17.9 Time-Interleaved A/D Converters 687
17.10 Key Points 690
17.11 References 691
17.12 Problems 692

CHAPTER 18 OVERSAMPLING CONVERTERS 696

18.1 Oversampling without Noise Shaping 696
18.1.1 Quantization Noise Modelling 697
18.1.2 White Noise Assumption 697
18.1.3 Oversampling Advantage 699
18.1.4 The Advantage of 1-Bit D/A Converters 701
18.2 Oversampling with Noise Shaping 702
18.2.1 Noise-Shaped Delta-Sigma Modulator 703
18.2.2 First-Order Noise Shaping 704
18.2.3 Switched-Capacitor Realization of a First-Order A/D Converter 706
18.2.4 Second-Order Noise Shaping 706
18.2.5 Noise Transfer-Function Curves 708
18.2.6 Quantization Noise Power of 1-Bit Modulators 709
18.2.7 Error-Feedback Structure 709
18.3 System Architectures 711
18.3.1 System Architecture of Delta-Sigma A/D Converters 711
18.3.2 System Architecture of Delta-Sigma D/A Converters 713
18.4 Digital Decimation Filters 714
18.4.1 Multi-Stage 715
18.4.2 Single Stage 717
18.5 Higher-Order Modulators 718
18.5.1 Interpolative Architecture 718
18.5.2 Multi-Stage Noise Shaping (MASH) Architecture 719
18.6 Bandpass Oversampling Converters 721
18.7 Practical Considerations 722
18.7.1 Stability 722
18.7.2 Linearity of Two-Level Converters 723
18.7.3 Idle Tones 725
18.7.4 Dithering 726
18.7.5 Opamp Gain 726
18.8 Multi-Bit Oversampling Converters 727
18.8.1 Dynamic Element Matching 727
18.8.2 Dynamically Matched Current Source D/S Converters 728
18.8.3 Digital Calibration A/D Converter 728
18.8.4 A/D with Both Multi-Bit and Single-Bit Feedback 729
18.9 Third-Order A/D Design Example 730
18.10 Key Points 732
18.11 References 734
18.12 Problems 735

CHAPTER 19 PHASE-LOCKED LOOPS 738

19.1 Basic Phase-Locked Loop Architecture 738
19.1.1 Voltage Controlled Oscillator 739
19.1.2 Divider 740
19.1.3 Phase Detector 741
19.1.4 Loop Filer 746
19.1.5 The PLL in Lock 747
19.2 Linearized Small-Signal Analysis 748
19.2.1 Second-Order PLL Model 749
19.2.2 Limitations of the Second-Order Small-Signal Model 751
19.2.3 PLL Design Example 754
19.3 Jitter and Phase Noise 756
19.3.1 Period Jitter 760
19.3.2 P-Cycle Jitter 761
19.3.3 Adjacent Period Jitter 761
19.3.4 Other Spectral Representations of Jitter 762
19.3.5 Probability Density Function of Jitter 764
19.4 Electronic Oscillators 765
19.4.1 Ring Oscillators 766
19.4.2 LC Oscillators 771
19.4.3 Phase Noise of Oscillators 772
19.5 Jitter and Phase Noise in PLLS 777
19.5.1 Input Phase Noise and Divider Phase Noise 777
19.5.2 VCO Phase Noise 778
19.5.3 Loop Filter Noise 779
19.6 Key Points 781
19.7 References 782
19.8 Problems 782

INDEX 787

CHAPTER

1 Integrated-Circuit Devices and Modelling

In this chapter, both the operation and modelling of semiconductor devices are described. Although it is possible to do simple integrated-circuit design with a basic knowledge of semiconductor device modelling, for state-of-the-art design, an in-depth understanding of the second-order effects of device operation and their modelling is considered critical.

It is assumed that most readers have been introduced to transistors and their basic modelling in a previous course. Thus, fundamental semiconductor concepts are only briefly reviewed. Section 1.1 describes pn junctions (or diodes). This section is important in understanding the parasitic capacitances in many device models, such as junction capacitances. Section 1.2 covers the basics of MOS transistors and modelling. A summary of device models and important equations is presented in Section 1.3. This summary is particularly useful for a reader who already has a good background in transistor modelling, in which case the summary can be used to follow the notation used throughout the remainder of this book. Advanced MOS modelling is treated in Section 1.4, including behavior not covered by a simple square-law voltage–current relationship. It should be noted that all sections on MOS devices rely to some degree on the material previously presented in Section 1.1, in which depletion capacitance is covered. In addition, a brief description is given of the most important process-related parameters used in SPICE modelling in Section 1.5. Since passive devices are often required for analog design, the most common passives on integrated circuits are described in Section 1.6. Finally, this chapter concludes with an Appendix containing derivations of the more physically-based device equations.

1.1 SEMICONDUCTORS AND pn JUNCTIONS

A semiconductor is a crystal lattice structure that can have free electrons (which are negative carriers) and/or free holes (which are an absence of electrons and are equivalent to positive carriers). The type of semiconductor typically used is silicon, an abundant element found, for example, in high concentrations in sand. This material has a valence of four, implying that each atom has four electrons to share with neighboring atoms when forming the covalent bonds of the crystal lattice. *Intrinsic* silicon (i.e., undoped silicon) is a very pure crystal structure that has equal numbers of free electrons and holes. These free carriers are those electrons that have gained enough energy due to thermal agitation to escape their bonds, and the resulting holes that they leave behind. At room temperature, there are approximately 1.1×10^{10} carriers of each type per cm^3, or equivalently $n_i = 1.1 \times 10^{16}$ carriers/m^3, defined as the carrier concentration of intrinsic silicon. The number of carriers approximately doubles for every 11 °C increase in temperature.

If one dopes silicon with a pentavalent impurity (i.e., atoms of an element having a valence of five, or equivalently five electrons in the outer shell, available when bonding with neighboring atoms), there will be almost one extra free electron for every impurity atom.[1] These free electrons can be used to conduct current. A pentavalent

1. In fact, there will be slightly fewer mobile carriers than the number of impurity atoms since some of the free electrons from the dopants have recombined with holes. However, since the number of holes of intrinsic silicon is much less than typical doping concentrations, this inaccuracy is small.

impurity is said to *donate* free electrons to the silicon crystal, and thus the impurity is known as a *donor*. Examples of donor elements are phosphorus, P, and arsenic, As. These impurities are also called n-type dopants since the free carriers resulting from their use have negative charge. When an n-type impurity is used, the total number of negative carriers or electrons is almost the same as the doping concentration, and is much greater than the number of free electrons in intrinsic silicon. In other words,

$$n_n = N_D \tag{1.1}$$

where n_n denotes the free-electron concentration in n-type material and N_D is the doping concentration (with the subscript D denoting donor). On the other hand, the number of free holes in n-doped material will be much less than the number of holes in intrinsic silicon and can be shown [Sze, 1981] to be given by

$$p_n = \frac{n_i^2}{N_D} \tag{1.2}$$

Similarly, if one dopes silicon with atoms that have a valence of three, for example, boron (B), the concentration of positive carriers or holes will be approximately equal to the *acceptor* concentration, N_A,

$$p_p = N_A \tag{1.3}$$

and the number of negative carriers in the p-type silicon, n_p, is given by

$$n_p = \frac{n_i^2}{N_A} \tag{1.4}$$

EXAMPLE 1.1

Intrinsic silicon is doped with boron at a concentration of 10^{26} atoms/m^3. At room temperature, what are the concentrations of holes and electrons in the resulting doped silicon? Assume that $n_i = 1.1 \times 10^{16}$ carriers/m^3.

Solution

The hole concentration, p_p, will approximately equal the doping concentration ($p_p = N_A = 10^{26}$ holes/m^3). The electron concentration is found from (1.4) to be

$$n_p = \frac{(1.1 \times 10^{16})^2}{10^{26}} = 1.2 \times 10^6 \text{ electrons/m}^3 \tag{1.5}$$

Such doped silicon is referred to as p-type since it has many more free holes than free electrons.

1.1.1 Diodes

To realize a diode, also called a pn junction, one part of a semiconductor is doped n-type and an adjacent part is doped p-type, as shown in Fig. 1.1. The diode, or junction, is formed between the p^+ region and the n region, where superscripts are used to indicate the relative doping levels. For example, the p^- bulk region in Fig. 1.1 might have an impurity concentration of 5×10^{21} carriers/m^3, whereas the p^+ and n^+ regions would be doped more

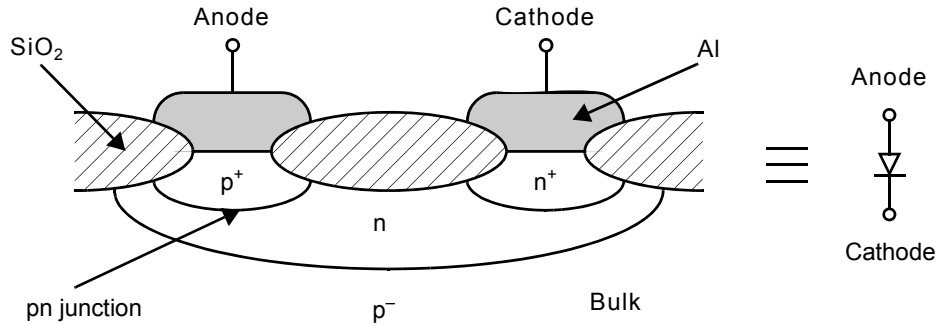


Fig. 1.1 A cross section of a pn diode.

heavily to a value around 10^{25} to 10^{27} carriers/m^3.[2] Also, note that the metal contacts to the diode (in this case, aluminum) are connected to heavily doped regions, otherwise a *Schottky diode* would form. (Schottky diodes are discussed on page 13.) Thus, in order not to make a Schottky diode, the connection to the n region is actually made via the n^+ region.

In the p^+ side, a large number of free positive carriers are available, whereas in the n side, many free negative carriers are available. The holes in the p^+ side will tend to disperse or diffuse into the n side, whereas the free electrons in the n side will tend to diffuse to the p^+ side. This process is very similar to two gases randomly diffusing together. This diffusion lowers the concentration of free carriers in the region between the two sides. As the two types of carriers diffuse together, they recombine. Every electron that diffuses from the n side to the p side leaves behind a *bound* positive charge close to the transition region. Similarly, every hole that diffuses from the p side leaves behind a bound electron near the transition region. The end result is shown in Fig. 1.2. This diffusion of free carriers creates a *depletion region* at the junction of the two sides where no free carriers exist, and which has a net negative charge on the p^+ side and a net positive charge on the n side. The total amount of exposed or bound charge on the two sides of the junction must be equal for charge neutrality. This requirement causes the depletion region to extend farther into the more lightly doped n side than into the p^+ side.

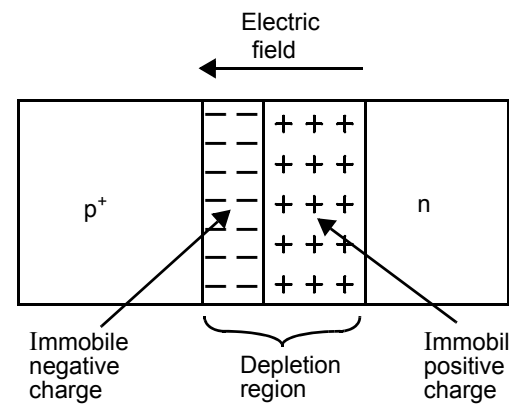


Fig. 1.2 A simplified model of a diode. Note that a depletion region exists at the junction due to diffusion and extends farther into the more lightly doped side.

As these bound charges are exposed, an electric field develops going from the n side to the p^+ side. This electric field gives rise to a potential difference between the n and p^+ sides, called the built-in voltage of the junction. It opposes the diffusion of free carriers until there is no net movement of charge under open-circuit and steady-state conditions. The built-in voltage of an open-circuit pn junction is [Sze 1981]

$$\Phi_0 = V_T \ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{1.6}$$

2. For reference, there are roughly 5×10^{28} atoms/m^3 in pure crystalline silicon.

where

$$V_T = \frac{kT}{q} \tag{1.7}$$

T is the temperature in degrees Kelvin ($\cong$ 300 °K at room temperature), k is Boltzmann's constant (1.38×10^{-23} JK^{-1}), and q is the charge of an electron (1.602×10^{-19} C). At room temperature, V_T is approximately 26 mV.

EXAMPLE 1.2

A pn junction has $N_A = 10^{25}$ holes/m^3 and $N_D = 10^{22}$ electrons/m^3. What is the built-in junction potential? Assume that $n_i = 1.1 \times 10^{16}$ carriers/m^3.

Solution

Using (1.6), we obtain

$$\Phi_0 = 0.026 \times \ln\left(\frac{10^{25} \times 10^{22}}{(1.1 \times 10^{16})^2}\right) = 0.89 \text{ V} \tag{1.8}$$

This is a typical value for the built-in potential of a junction with one side heavily doped. As an approximation, we will normally use $\Phi_0 \cong 0.9$ V for the built-in potential of a junction having one side heavily doped.

1.1.2 Reverse-Biased Diodes

A silicon diode having an anode-to-cathode (i.e., p side to n side) voltage of 0.4 V or less will not be conducting appreciable current. In this case, it is said to be *reverse-biased.* If a diode is reverse-biased, current flow is primarily due to thermally generated carriers in the depletion region, and it is extremely small. Although this reverse-biased current is only weakly dependent on the applied voltage, *the reverse-biased current is directly proportional to the area of the diode junction.* However, an effect that should not be ignored, particularly at high frequencies, is the junction capacitance of a diode. In reverse-biased diodes, this junction capacitance is due to varying charge storage in the depletion regions and is modelled as a *depletion capacitance.*

To determine the depletion capacitance, we first state the relationship between the depletion widths and the applied reverse voltage, V_R [Sze, 1981].

$$x_n = \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)}{q}\frac{N_A}{N_D(N_A + N_D)}\right]^{1/2} \tag{1.9}$$

$$x_p = \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)}{q}\frac{N_D}{N_A(N_A + N_D)}\right]^{1/2} \tag{1.10}$$

Here, ε_0 is the permittivity of free space (equal to 8.854×10^{-12} F/m), V_R is the reverse-bias voltage of the diode, and K_s is the relative permittivity of silicon (equal to 11.8). These equations assume an abrupt junction where the doping changes instantly from the n to the p side. Modifications to these equations for graded junctions are treated in the next section.

From the above equations, we see that if one side of the junction is more heavily doped than the other, the depletion region will extend mostly on the lightly doped side. For example, if $N_A >> N_D$ (i.e., if the p region is more heavily doped), we can approximate (1.9) and (1.10) as

$$x_n \cong \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)}{qN_D}\right]^{1/2} \quad x_p \cong \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)N_D}{qN_A^2}\right]^{1/2} \tag{1.11}$$

Indeed, for this case

$$\frac{x_n}{x_p} \cong \frac{N_A}{N_D} \tag{1.12}$$

This special case is called a *single-sided diode*.

EXAMPLE 1.3

For a pn junction having $N_A = 10^{25}$ holes/m^3 and $N_D = 10^{22}$ electrons/m^3, what are the depletion region depths for a 1-V reverse-bias voltage?

Solution

Since $N_A >> N_D$ and we already have found in Example 1.2 that $\Phi_0 = 0.9$ V, we can use (1.11) to find

$$x_n = \left[\frac{2 \times 11.8 \times 8.854 \times 10^{-12} \times 1.9}{1.6 \times 10^{-19} \times 10^{22}}\right]^{1/2} = 0.50\ \mu\text{m} \tag{1.13}$$

$$x_p = \frac{x_n}{(N_A/N_D)} = 0.50\ \text{nm} \tag{1.14}$$

Note that the depletion region width in the lightly doped n region is 1,000 times greater than that in the more heavily doped p region.

The charge stored in the depletion region, per unit cross-sectional area, is found by multiplying the depletion region width by the concentration of the immobile charge (which is approximately equal to q times the impurity doping density). For example, on the n side, we find the charge in the depletion region to be given by multiplying (1.9) by qN_D, resulting in

$$Q^+ = \left[2qK_s\varepsilon_0(\Phi_0 + V_R)\frac{N_AN_D}{N_A + N_D}\right]^{1/2} \tag{1.15}$$

This amount of charge must also equal Q^- on the p side since there is charge equality. In the case of a single-sided diode when $N_A >> N_D$, we have

$$Q^- = Q^+ \cong [2qK_s\varepsilon_0(\Phi_0 + V_R)N_D]^{1/2} \tag{1.16}$$

Key Point: *The charge–voltage relationship of a reverse-biased pn junction is modeled by a nonlinear depletion capacitance.*

Note that this result is independent of the impurity concentration on the heavily doped side. Thus, we see from the above relation that the charge stored in the depletion region is nonlinearly dependent on the applied reverse-bias voltage. This charge–voltage relationship is modelled by a nonlinear depletion capacitance.

For small changes in the reverse-biased junction voltage, about a bias voltage, we can find an equivalent *small-signal* capacitance, C_j, by differentiating (1.15) with respect to V_R. Such a differentiation results in

$$C_j = \frac{dQ^+}{dV_R} = \left[\frac{qK_s\varepsilon_0}{2(\Phi_0 + V_R)}\frac{N_A N_D}{N_A + N_D}\right]^{1/2} = \frac{C_{j0}}{\sqrt{1 + \dfrac{V_R}{\Phi_0}}} \tag{1.17}$$

where C_{j0} is the depletion capacitance per unit area at $V_R = 0$ and is given by

$$C_{j0} = \sqrt{\frac{qK_s\varepsilon_0}{2\Phi_0}\frac{N_A N_D}{N_A + N_D}} \tag{1.18}$$

In the case of a one-sided diode with $N_A >> N_D$, we have

$$C_j = \left[\frac{qK_s\varepsilon_0 N_D}{2(\Phi_0 + V_R)}\right]^{1/2} = \frac{C_{j0}}{\sqrt{1 + \dfrac{V_R}{\Phi_0}}} \tag{1.19}$$

where now

$$C_{j0} = \sqrt{\frac{qK_s\varepsilon_0 N_D}{2\Phi_0}} \tag{1.20}$$

It should be noted that many of the junctions encountered in integrated circuits are one-sided junctions with the lightly doped side being the substrate or sometimes what is called the *well.* The more heavily doped side is often used to form a contact to interconnecting metal. From (1.20), we see that, for these one-sided junctions, the depletion capacitance is approximately independent of the doping concentration on the heavily doped side, and is proportional to the square root of the doping concentration of the more lightly doped side. Thus, smaller depletion capacitances are obtained for more lightly doped substrates—a strong incentive to strive for lightly doped substrates.

Finally, note that by combining (1.15) and (1.18), we can express the equation for the immobile charge on either side of a reverse-biased junction as

$$Q = 2C_{j0}\Phi_0\sqrt{1 + \frac{V_R}{\Phi_0}} \tag{1.21}$$

As seen in Example 1.6, this equation is useful when one is approximating the large-signal charging (or discharging) time for a reverse-biased diode.

EXAMPLE 1.4

For a pn junction having $N_A = 10^{25}$ holes/m^3 and $N_D = 10^{22}$ electrons/m^3, what is the total zero-bias depletion capacitance for a diode of area 10 μm $\times$ 10 μm? What is its depletion capacitance for a 3-V reverse-bias voltage?

Solution

Making use of (1.20), we have

$$C_{j0} = \sqrt{\frac{1.6 \times 10^{-19} \times 11.8 \times 8.854 \times 10^{-12} \times 10^{22}}{2 \times 0.9}} = 304.7 \ \mu\text{F/m}^2 \tag{1.22}$$

Since the diode area is 100×10^{-12} m^2, the total zero-bias depletion capacitance is

$$C_{T\text{-}j0} = 100 \times 10^{-12} \times 304.7 \times 10^{-6} = 30.5 \text{ fF} \tag{1.23}$$

At a 3-V reverse-bias voltage, we have from (1.19)

$$C_{T\text{-}j} = \frac{30.5 \text{ fF}}{\sqrt{1 + \left(\frac{3}{0.9}\right)}} = 14.7 \text{ fF} \tag{1.24}$$

We see a decrease in junction capacitance as the width of the depletion region is increased.

1.1.3 Graded Junctions

All of the above equations assumed an abrupt junction where the doping concentration changes quickly from p to n over a small distance. Although this is a good approximation for many integrated circuits, it is not always so. For example, the collector-to-base junction of a bipolar transistor is most commonly realized as a *graded* junction. In the case of graded junctions, the exponent 1/2 in (1.15) is inaccurate, and an exponent closer to unity is more accurate, perhaps 0.6 to 0.7. Thus, for graded junctions, (1.15) is typically written as

$$Q = \left[2qK_s\varepsilon_0(\Phi_0 + V_R)\frac{N_A N_D}{N_A + N_D}\right]^{1-m_j} \tag{1.25}$$

where m_j is a constant that depends upon the doping profile. For example, a linearly graded junction has $m_j = 1/3$.

Differentiating (1.25) to find the depletion capacitance, we have

$$C_j = (1 - m_j)\left[2qK_s\varepsilon_0\frac{N_A N_D}{N_A + N_D}\right]^{1-m_j}\frac{1}{(\Phi_0 + V_R)^{m_j}} \tag{1.26}$$

This depletion capacitance can also be written as

$$C_j = \frac{C_{j0}}{\left(1 + \frac{V_R}{\Phi_0}\right)^{m_j}} \tag{1.27}$$

where

$$C_{j0} = (1 - m_j)\left[2qK_s\varepsilon_0\frac{N_A N_D}{N_A + N_D}\right]^{1-m_j}\frac{1}{\Phi_0^{m_j}} \tag{1.28}$$

From (1.27), we see that a graded junction results in a depletion capacitance that is less dependent on V_R than the equivalent capacitance in an abrupt junction. In other words, since m is less than 0.5, the depletion capacitance for a graded junction is more linear than that for an abrupt junction. Correspondingly, increasing the reverse-bias voltage for a graded junction is not as effective in reducing the depletion capacitance as it is for an abrupt junction.

Finally, as in the case of an abrupt junction, the depletion charge on either side of the junction can also be written as

$$Q = \frac{C_{j0}}{1 - m_j}\Phi_0\left(1 + \frac{V_R}{\Phi_0}\right)^{1-m_j} \tag{1.29}$$

EXAMPLE 1.5

Repeat Example 1.4 for a graded junction with $m_j = 0.4$.

Solution

Noting once again that $N_A >> N_D$, we approximate (1.28) as

$$C_{j0} = (1 - m_j)[2qK_s\varepsilon_0 N_D]^{1-m_j}\frac{1}{\Phi_0^{m_j}} \tag{1.30}$$

resulting in

$$C_{j0} = 81.5\ \mu F/m^2 \tag{1.31}$$

which, when multiplied by the diode's area of $10\ \mu m \times 10\ \mu m$, results in

$$C_{T\text{-}j0} = 8.1\ fF \tag{1.32}$$

For a 3-V reverse-bias voltage, we have

$$C_{T\text{-}j} = \frac{8.1 fF}{(1 + 3/0.9)^{0.4}} = 4.5\ fF \tag{1.33}$$

1.1.4 Large-Signal Junction Capacitance

The equations for the junction capacitance given above are only valid for small changes in the reverse-bias voltage. This limitation is due to the fact that C_j depends on the size of the reverse-bias voltage instead of being a constant. As a result, it is extremely difficult and time consuming to accurately take this nonlinear capacitance into account when calculating the time to charge or discharge a junction over a large voltage change. A commonly used approximation when analyzing the transient response for large voltage changes is to use an *average size* for the junction capacitance by calculating the junction capacitance at the two extremes of the reverse-bias voltage. Unfortunately, a problem with this approach is that when the diode is forward biased with $V_R \cong -\Phi_0$, equation (1.17) "blows up" (i.e., is equal to infinity). To circumvent this problem, one can instead calculate the charge stored in the junction for the two extreme values of applied voltage (through the use of (1.21)), and then through the use of $Q = CV$, calculate the average capacitance according to

$$C_{j\text{-}av} = \frac{Q(V_2) - Q(V_1)}{V_2 - V_1} \tag{1.34}$$

where V_1 and V_2 are the two voltage extremes [Hodges, 1988].

From (1.21), for an abrupt junction with reverse-bias voltage V_i, we have

$$Q(V_i) = 2C_{j0}\Phi_0\sqrt{1 + \frac{V_i}{\Phi_0}} \tag{1.35}$$

Therefore,

$$C_{j\text{-}av} = 2C_{j0}\Phi_0\frac{\left(\sqrt{1 + \frac{V_2}{\Phi_0}} - \sqrt{1 + \frac{V_1}{\Phi_0}}\right)}{V_2 - V_1} \tag{1.36}$$

EXAMPLE 1.6

For the circuit shown in Fig. 1.3, where a reverse-biased diode is being charged from 0 V to 1 V, through a 10-kΩ resistor, calculate the time required to charge the diode from 0 V to 0.7 V. Assume that $C_{j0} = 0.2$ fF/(μm)^2 and that the diode has an area of 20 μm × 5 μm. Compare your answer to that obtained using SPICE. Repeat the question for the case of the diode being discharged from 1 V to 0.3 V.

Solution

For the special case of $V_1 = 0V$ and $V_2 = 1V$, and using $\Phi_0 = 0.9$ V in equation (1.36) we find that

$$C_{j\text{-}av} = 0.815C_{j0} \tag{1.37}$$

Thus, as a rough approximation to quickly estimate the charging time of a junction capacitance from 0 V to 1 V (or vice versa), one can use

$$C_{j\text{-}av} \cong 0.8C_{j0} \tag{1.38}$$

The total small-signal capacitance of the junction at 0-V bias voltage is obtained by multiplying 0.2 fF/(μm)^2 by the junction area to obtain

$$C_{T\text{-}j0} = 0.2 \times 10^{-15} \times 20 \times 5 = 0.02 \text{ pF} \tag{1.39}$$

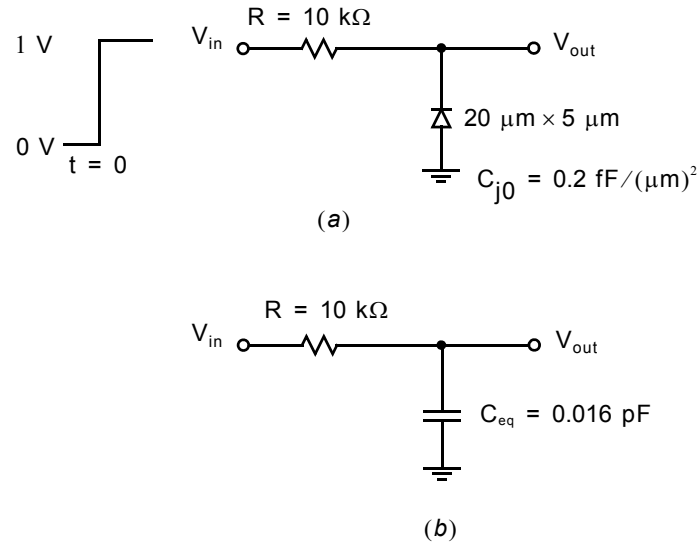


Fig. 1.3 (*a*) The circuit used in Example 1.6; (*b*) its RC approximate equivalent.

Using (1.37), we have

$$C_{T\text{-}j\text{-}av} = 0.815 \times 0.02 = 0.016 \text{ pF} \tag{1.40}$$

resulting in a time constant of

$$\tau = RC_{j\text{-}av} = 0.16 \text{ ns} \tag{1.41}$$

It is not difficult to show that the time it takes for a first-order circuit to rise (or fall) 70 percent of its final value is equal to 1.2τ. Thus, in this case,

$$t_{70\%} = 1.2\tau = 0.20 \text{ ns} \tag{1.42}$$

SPICE! *Refer to the book web site for a netlist.*

As a check, the circuit of Fig. 1.3(*a*) was analyzed using SPICE. The SPICE simulation gave a 0-V to 0.7-V rise time of 0.21 ns and a 1-V to 0.3-V fall time of 0.19 ns, in general agreement with the 0.20 ns predicted. The reason for the different values of the rise and fall times is the nonlinearity of the junction capacitance. For smaller bias voltages it is larger than that predicted by (1.37), whereas for larger bias voltages it is smaller. Normally, the extra accuracy that results from performing a more accurate analysis is not worth the extra complication because one seldom knows the value of C_{j0} to better than 20 percent accuracy.

1.1.5 Forward-Biased Junctions

A positive voltage applied from the p side to the n side of a diode reduces the electric field opposing the diffusion of the free carriers across the depletion region. It also reduces the width of the depletion region. If this forward-bias voltage is large enough, the carriers will start to diffuse across the junction, resulting in a current flow from the anode to the cathode. For silicon, appreciable diode current starts to occur for a forward-bias

voltage around 0.5 V. For germanium and gallium arsenide semiconductor materials, current conduction starts to occur around 0.3 V and 0.9 V, respectively.

When the junction potential is sufficiently lowered for conduction to occur, the carriers diffuse across the junction due to the large gradient in the mobile carrier concentrations. Note that there are more carriers diffusing from the heavily doped side to the lightly doped side than from the lightly doped side to the heavily doped side.

After the carriers cross the depletion region, they greatly increase the *minority charge* at the edge of the depletion region. These minority carriers will diffuse away from the junction toward the bulk. As they diffuse, they recombine with the majority carriers, thereby decreasing their concentration. This concentration gradient of the minority charge (which decreases the farther one gets from the junction) is responsible for the current flow near the junction.

The majority carriers that recombine with the diffusing minority carriers come from the metal contacts at the junctions because of the forward-bias voltage. These majority carriers flow across the bulk, from the contacts to the junction, due to an electric field applied across the bulk. This current flow is called *drift*. It results in small potential drops across the bulk, especially in the lightly doped side. Typical values of this voltage drop might be 50 mV to 0.1 V, depending primarily on the doping concentration of the lightly doped side, the distance from the contacts to the junction, and the cross-sectional area of the junction.

In the forward-bias region, the current–voltage relationship is exponential and can be shown (see Appendix) to be

$$I_D = I_S e^{V_D/V_T} \tag{1.43}$$

where V_D is the voltage applied across the diode and

$$I_S \propto A_D\left(\frac{1}{N_A} + \frac{1}{N_D}\right) \tag{1.44}$$

I_S is known as the *scale current* and is seen to be proportional to the area of the diode junction, A_D, and inversely proportional to the doping concentrations.

1.1.6 Junction Capacitance of Forward-Biased Diode

When a junction changes from reverse biased (with little current through it) to forward biased (with significant current flow across it), the charge being stored near and across the junction changes. Part of the change in charge is due to the change in the width of the depletion region and therefore the amount of immobile charge stored in it. This change in charge is modelled by the depletion capacitance, C_j, similar to when the junction is reverse biased. An additional change in charge storage is necessary to account for the change of the minority carrier concentration close to the junction required for the diffusion current to exist. For example, if a forward-biased diode current is to double, then the slopes of the minority charge storage at the diode junction edges must double, and this, in turn, implies that the minority charge storage must double. This component is modelled by another capacitance, called the *diffusion capacitance,* and denoted C_d.

The diffusion capacitance can be shown (see Appendix) to be

$$C_d = \tau_T \frac{I_D}{V_T} \tag{1.45}$$

where τ_T is the transit time of the diode. Normally τ_T is specified for a given technology, so that one can calculate the diffusion capacitance. *Note that the diffusion capacitance of a forward-biased junction is proportional to the diode current.*

The total capacitance of the forward-biased junction is the sum of the diffusion capacitance, C_d, and the depletion capacitance, C_j. Thus, the total junction capacitance is given by

$$C_T = C_d + C_j \tag{1.46}$$

For a forward-biased junction, the depletion capacitance, C_j, can be roughly approximated by $2C_{j0}$. The accuracy of this approximation is not critical since the diffusion capacitance is typically much larger than the depletion capacitance.

Finally, it should be mentioned that as a diode is turned off for a short period of time a current will flow in the negative direction until the minority charge is removed. This behavior does not occur in Schottky diodes since they do not have minority charge storage.

1.1.7 Small-Signal Model of a Forward-Biased Diode

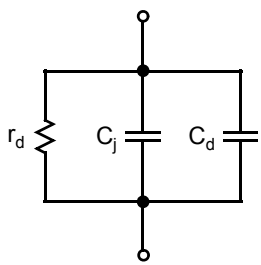


Fig. 1.4 The small-signal model for a forward-biased junction.

A small-signal equivalent model for a forward-biased diode is shown in Fig. 1.4. A resistor, r_d, models the change in the diode voltage, V_D, that occurs when I_D changes. Using (1.43), we have

$$\frac{1}{r_d} = \frac{dI_D}{dV_D} = I_S \frac{e^{V_D/V_T}}{V_T} = \frac{I_D}{V_T} \tag{1.47}$$

This resistance is called the incremental resistance of the diode. For very accurate modelling, it is sometimes necessary to add the series resistance due to the bulk and also the resistance associated with the contacts. Typical values for the contact resistance (caused by the work-function[3] difference between metal and silicon) might be $20\ \Omega$ to $40\ \Omega$.

By combining (1.45) and (1.47), we see that an alternative equation for the diffusion capacitance, C_d, is

$$C_d = \frac{\tau_T}{r_d} \tag{1.48}$$

Since for moderate forward-bias currents, $C_d >> C_j$, the total small-signal capacitance is $C_T \cong C_d$, and

$$r_d C_T \cong \tau_T \tag{1.49}$$

Thus, for charging or discharging a forward-biased junction with a current source having an impedance much larger than r_d, the time constant of the charging is approximately equal to the transit time of the diode and is independent of the diode current. For smaller diode currents, where C_j becomes important, the charging or discharging time constant of the circuit becomes larger than τ_T.

EXAMPLE 1.7

A given diode has a transit time of 100 ps and is biased at 1 mA. What are the values of its small-signal resistance and diffusion capacitance? Assume room temperature, so that $V_T = kT/q = 26$ mV.

3. The work-function of a material is defined as the minimum energy required to remove an electron at the Fermi level to the outside vacuum region.

Solution

We have

$$r_d = \frac{V_T}{I_D} = \frac{26\ \text{mV}}{1\ \text{mA}} = 26\ \Omega$$

and

$$C_d = \frac{\tau_T}{r_d} = \frac{100\ \text{ps}}{26\ \Omega} = 3.8\ \text{pF}$$

Note that this diffusion capacitance is over 100 times larger than the total depletion capacitance found in Examples 1.4 and 1.5.

1.1.8 Schottky Diodes

A different type of diode, one sometimes used in microcircuit design, is realized by contacting metal to a lightly doped semiconductor region (rather than a heavily doped region) as shown in Fig. 1.5. Notice that the aluminum anode is in direct contact with a relatively lightly doped n^- region. Because the n^- region is relatively lightly doped, the work-function difference between the aluminum contact and the n^- silicon is larger than would be the case for aluminum contacting to an n^+ region, as occurs at the cathode. This causes a depletion region and, correspondingly, a diode to occur at the interface between the aluminum anode and the n^- silicon region. This diode has different characteristics than a normal pn junction diode. First, its voltage drop when forward biased is smaller. This voltage drop is dependent on the metal used; for aluminum it might be around 0.5 V. More importantly, when the diode is forward biased, there is no minority-charge storage in the lightly doped n^- region. Thus, the small-signal model of a forward-biased Schottky diode has $C_d = 0$ (with reference to Fig. 1.4). The absence of this diffusion capacitance makes the diode much faster. It is particularly faster when turning off, because it is not necessary to remove the minority charge first. Rather, it is only necessary to discharge the depletion capacitance through about 0.2 V.

Schottky diodes have been used extensively in bipolar logic circuits. They are also used in a number of high-speed analog circuits, particularly those realized in gallium arsenide (GaAs) technologies, rather than silicon technologies.

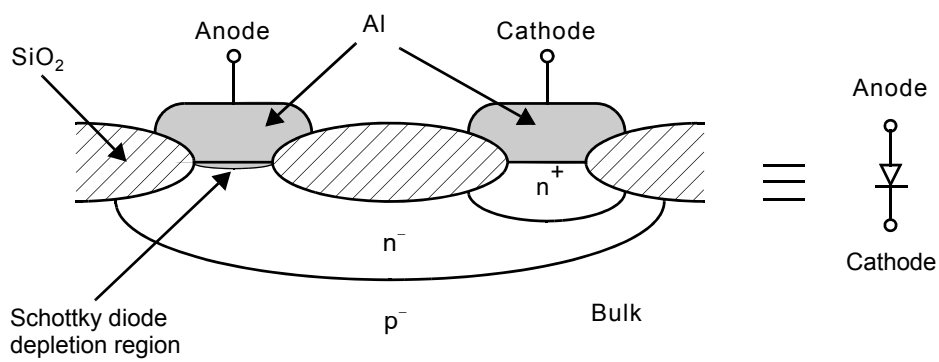


Fig. 1.5 A cross section of a Schottky diode.

1.2 MOS TRANSISTORS

Presently, the most popular technology for realizing microcircuits makes use of MOS transistors. Unlike most bipolar junction transistor (BJT) technologies, which make dominant use of only one type of transistor (npn transistors in the case of BJT processes[4]), MOS circuits normally use two complementary types of transistors—n-channel and p-channel. While n-channel devices conduct with a positive gate voltage, p-channel devices conduct with a negative gate voltage. Moreover, electrons are used to conduct current in n-channel transistors, while holes are used in p-channel transistors. Microcircuits containing both n-channel and p-channel transistors are called CMOS circuits, for *complementary MOS*. The acronym MOS stands for *metal-oxide semiconductor,* which historically denoted the gate, insulator, and channel region materials, respectively. However, most present CMOS technologies utilize polysilicon gates rather than metal gates.

Before CMOS technology became widely available, most MOS processes made use of only n-channel transistors (NMOS). However, often two different types of n-channel transistors could be realized. One type, enhancement n-channel transistors, is similar to the n-channel transistors realized in CMOS technologies. Enhancement transistors require a positive gate-to-source voltage to conduct current. The other type, depletion transistors, conduct current with a gate-source voltage of 0 V. Depletion transistors were used to create high-impedance loads in NMOS logic gates.

Key Point: *The source terminal of an n-channel transistor is defined as whichever of the two terminals has a lower voltage. For a p-channel transistor, the source would be the terminal with the higher voltage.*

A typical cross section of an n-channel enhancement-type MOS transistor is shown in Fig. 1.6. With no voltage applied to the gate, the n^+ source and drain regions are separated by the p^- substrate. The distance between the drain and the source is called the channel length, L. In present MOS technologies, the minimum channel length may be as small as 28 nm. It should be noted that there is no physical difference between the drain and the source.[5] The source terminal of an n-channel transistor is defined as whichever of the two terminals has a lower voltage. For a p-channel transistor, the source would be the terminal with the higher voltage. When a transistor is turned on, current flows from the drain to the source in an n-channel transistor and from the source to the drain in a p-channel transistor. In both cases, the true carriers travel from the source to drain, but the current directions are different because n-channel carriers (electrons) are negative, whereas p-channel carriers (holes) are positive.

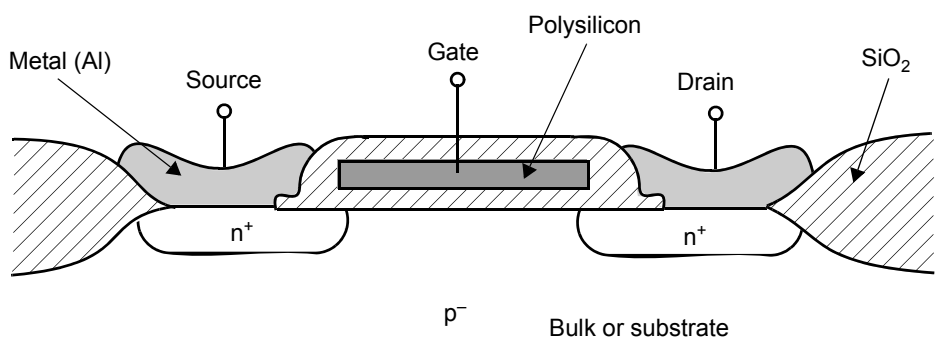


Fig. 1.6 A cross section of a typical n-channel transistor.

4. Most BJT technologies can also realize low-speed lateral pnp transistors. Normally these would only be used to realize current sources as they have low gains and poor frequency responses. Recently, bipolar technologies utilizing high-speed vertical pnp transistors, as well as high-speed npn transistors, have become available and are growing in popularity. These technologies are called complementary bipolar technologies.

5. Large MOS transistors used for power applications are an exception as they might not be realized with symmetric drain and source junctions.

The gate is normally realized using polysilicon, which is heavily doped noncrystalline (or amorphous) silicon. Polysilicon gates are used (instead of metal) because polysilicon has allowed the dimensions of the transistor to be realized much more accurately during the patterning of the transistor. This higher geometric accuracy has resulted in smaller, faster transistors. However, due to the relatively higher resistance of polysilicon, there are continuous efforts to realize metal gates in CMOS fabrication technologies.

The gate is physically separated from the surface of the silicon by a thin insulator made of silicon dioxide (SiO_2). Thus, the gate is electrically isolated from the channel and affects the channel (and hence, the transistor current) only through electrostatic (capacitive) coupling. The typical thickness of the SiO_2 insulator between the gate and the channel is presently between 1 to 30 nm. Since the gate is electrically isolated from the channel, it does not conduct appreciable dc current. However, because of the inherent capacitances in MOS transistors, transient gate currents do exist when gate voltage is quickly changing.

Normally the p^- substrate (or bulk) is connected to the most negative voltage in a microcircuit. In analog circuits, this might be the negative power supply, and in digital circuits it is normally ground or 0 V. This connection results in all transistors placed in the substrate being surrounded by reverse-biased junctions, which electrically isolate the transistors and thereby prevent conduction between the transistor terminals and the substrate (unless, of course, they are connected together through some other means).

1.2.1 Symbols for MOS Transistors

Many symbols have been used to represent MOS transistors. Figure 1.7 shows some of the symbols that have been used to represent n-channel MOS transistors. The symbol in Fig. 1.7(*a*) is often used; note that there is nothing in the symbol to specify whether the transistor is n-channel or p-channel. A common rule is to assume, when in doubt, that the transistor is an n-channel transistor. The symbol in Fig. 1.7(*a*) will be used occasionally in this text when there is no need to distinguish between the drain and source terminals. Figure 1.7(*b*) is the most commonly used symbol for an n-channel transistor in analog design and is used most often throughout this text. An arrow points outward on the source terminal to indicate that the transistor is n-channel and indicates the direction of current.

MOS transistors are actually four-terminal devices, with the substrate being the fourth terminal. In n-channel devices, the p^- substrate is normally connected to the most negative voltage in the microcircuit, whereas for p-channel devices, the n^- substrate is normally connected to the most positive voltage. In these cases the substrate connection is normally not shown in the symbol. However, for CMOS technologies, at least one of the two types of transistors will be formed in a *well* substrate that need not be connected to one of the power supply nodes. For example, an n-well process would form n-channel transistors in a p^- substrate encompassing the entire microcircuit, while the p-channel transistors would be formed in many separate n-well substrates. In this case, most of the n-well substrates would be connected to the most positive power supply, while some might be connected to other nodes in the circuit (often the well is connected to the source of a transistor that is not connected to the power supply). In these cases, the symbol shown in Fig. 1.7(*c*) can be used to show the substrate connection explicitly. Note that the arrow points from the p substrate region to the n-channel region, just like the arrow in the diode symbol which points from the p anode to the n cathode region. It should be noted that this case is not encountered often in digital circuits and is more common in analog circuits. Sometimes, in the interest of simplicity, the isolation of the gate is not explicitly shown, as is the case of the symbol of Fig. 1.7(*d*). This simple notation is more common for

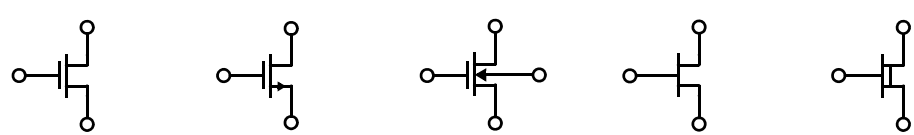


Fig. 1.7 Commonly used symbols for n-channel transistors.

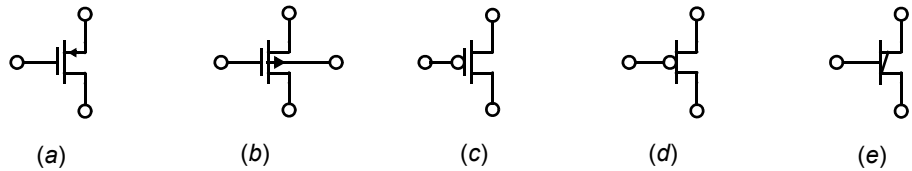


Fig. 1.8 Commonly used symbols for p-channel transistors.

digital circuits in which a large number of transistors are present. Since this symbol is also used for JFET transistors, it will never be used to represent MOS transistors in this text. The last symbol, shown in Fig. 1.7(*e*), denotes an n-channel depletion transistor. The extra line is used to indicate that a physical channel exists for a 0-V gate-source voltage. Depletion transistors were used in older NMOS technologies but are not typically available in CMOS processes.

Figure 1.8 shows some commonly used symbols for p-channel transistors. In this text, the symbol of Fig. 1.8(*a*) will be most often used. The symbol in Fig. 1.8(*c*) is sometimes used in digital circuits, where the circle indicates that a low voltage on the gate turns the transistor on, as opposed to a high voltage for an n-channel transistor Fig. 1.7(*a*). The symbols of Fig. 1.8(*d*) or Fig. 1.8(*e*) might be used in larger circuits where many transistors are present, to simplify the drawing somewhat. They will not be used in this text.

1.2.2 Basic Operation

The basic operation of MOS transistors will be described with respect to an n-channel transistor. First, consider the simplified cross sections shown in Fig. 1.9, where the source, drain, and substrate are all connected to ground. In this case, the MOS transistor operation is similar to a capacitor. The gate acts as one plate of the capacitor, and the surface of the silicon, just under the thin insulating SiO_2, acts as the other plate.

If the gate voltage is very negative, as shown in Fig. 1.9(*a*), positive charge will be attracted to the channel region. Since the substrate was originally doped p^-, this negative gate voltage has the effect of simply increasing the channel doping to p^+, resulting in what is called an *accumulated channel.* The n^+ source and drain regions are separated from the p^+-channel region by depletion regions, resulting in the equivalent circuit of two back-to-back diodes. Thus, only leakage current will flow even if one of the source or drain voltages becomes large (unless the drain voltage becomes so large as to cause the transistor to break down).

In the case of a positive voltage being applied to the gate, the opposite situation occurs, as shown in Fig. 1.9(*b*). For small positive gate voltages, the positive carriers in the channel under the gate are initially repulsed and the channel changes from a p^- doping level to a depletion region. As a more positive gate voltage is applied, the gate attracts negative charge from the source and drain regions, and the channel becomes an n region with mobile electrons connecting the drain and source regions.[6] In short, a sufficiently large positive gate-source voltage changes the channel beneath the gate to an n region, and the channel is said to be *inverted*.

The gate-source voltage, for which the concentration of electrons under the gate is equal to the concentration of holes in the p^- substrate far from the gate, is commonly referred to as the *transistor threshold voltage* and denoted V_{tn} (for n-channel transistors). For gate-source voltages larger than V_{tn}, there is an n-type channel present, and conduction between the drain and the source can occur. For gate-source voltages less than V_{tn}, it is normally assumed that the transistor is off and no current flows between the drain and the source. However, it should be noted that this assumption of zero drain-source current for a transistor that is

6. The drain and source regions are sometimes called diffusion regions or junctions for historical reasons. This use of the word *junction* is not synonymous with our previous use, in which it designated a pn interface of a diode.

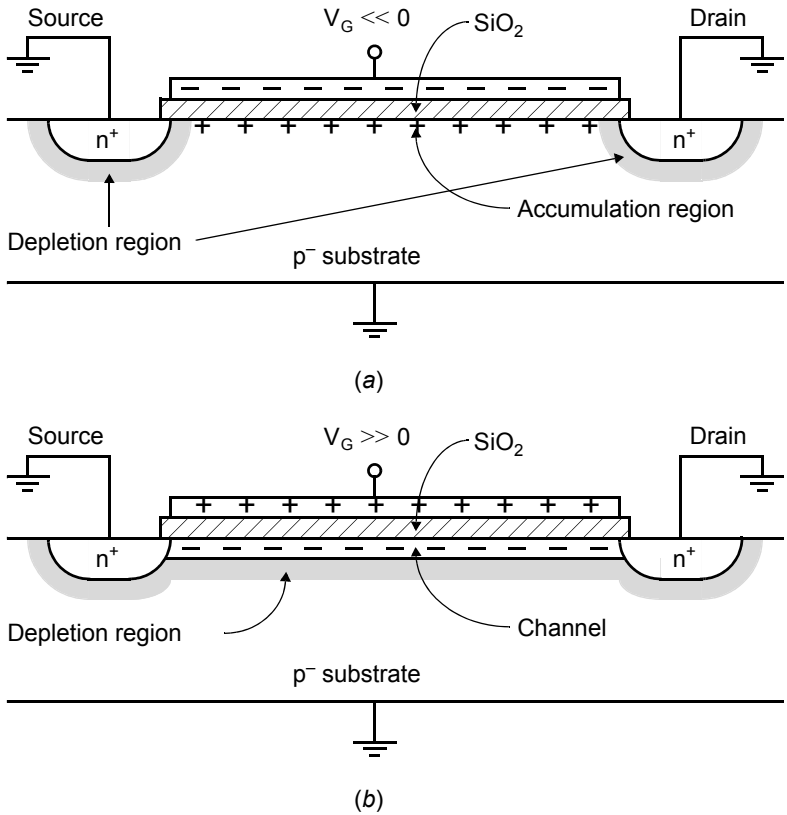


Fig. 1.9 An n-channel MOS transistor. (*a*) $V_G << 0$, resulting in an accumulated channel (no current flow); (*b*) $V_G >> 0$, and the channel is present (current flow possible from drain to source).

off is only an approximation. In fact, for gate voltages around V_{tn}, there is no abrupt current change, and for gate-source voltages slightly less than V_{tn}, small amounts of *subthreshold current* can flow, as discussed in Section 1.4.1.

When the gate-source voltage, V_{GS}, is larger than V_{tn}, the channel is present. As V_{GS} is increased, the density of electrons in the channel increases. Indeed, the carrier density, and therefore the charge density, is proportional to $V_{GS} - V_{tn}$, which is often called the *effective gate-source voltage* and denoted V_{eff}. Specifically, define

$$V_{eff} \equiv V_{GS} - V_{tn} \tag{1.50}$$

The charge density of electrons is then given by

$$Q_n = C_{ox}(V_{GS} - V_{tn}) = C_{ox}V_{eff} \tag{1.51}$$

Here, C_{ox} is the gate capacitance per unit area and is given by

$$C_{ox} = \frac{K_{ox}\varepsilon_0}{t_{ox}} \tag{1.52}$$

where K_{ox} is the relative permittivity of SiO_2 (approximately 3.9) and t_{ox} is the thickness of the thin oxide under the gate. A point to note here is that (1.51) is only accurate when both the drain and the source voltages are zero.

To obtain the total gate capacitance, (1.52) should be multiplied by the effective gate area, WL, where W is the gate width and L is the effective gate length. These dimensions are shown in Fig. 1.10. Thus the total gate capacitance, C_g, is given by

$$C_g = WLC_{ox} \tag{1.53}$$

and the total charge of the channel, $Q_{T\text{-}n}$, is given by

$$Q_{T\text{-}n} = WLC_{ox}(V_{GS} - V_{tn}) = WLC_{ox}V_{eff} \tag{1.54}$$

The gate capacitance is one of the major load capacitances that circuits must be capable of driving. Gate capacitances are also important when one is calculating *charge injection,* which occurs when a MOS transistor is being turned off because the channel charge, $Q_{T\text{-}n}$, must flow from under the gate out through the terminals to other places in the circuit.

Next, if the drain voltage is increased above 0 V, a drain-source potential difference exists. This difference results in current flowing from the drain to the source.[7] The relationship between V_{DS} and the drain-source current, I_D, is the same as for a resistor, assuming V_{DS} is small. This relationship is given [Sze, 1981] by

$$I_D = \mu_n Q_n \frac{W}{L} V_{DS} \tag{1.55}$$

where μ_n is the mobility of electrons near the silicon surface, and Q_n is the charge concentration of the channel per unit area (looking from the top down). Electron mobility is 0.14 m^2/Vs in pure intrinsic silicon, decreasing with increasing dopant concentrations to $\mu_n \cong 0.01 - 0.06$ m^2/Vs in modern NMOS devices. Note that as the channel length increases, the drain-source current decreases, whereas this current increases as either the charge density or the transistor width increases. Using (1.54) and (1.55) results in

$$I_D = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{tn})V_{DS} = \mu_n C_{ox} \frac{W}{L} V_{eff} V_{DS} \tag{1.56}$$

where it should be emphasized that this relationship is only valid for drain-source voltages near zero (i.e., V_{DS} much smaller than V_{eff}).

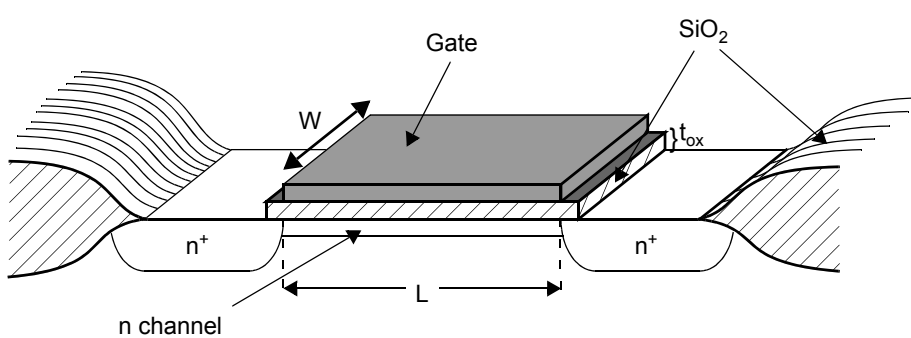


Fig. 1.10 The important dimensions of a MOS transistor.

7. The current is actually conducted by negative carriers (electrons) flowing from the source to the drain. Negative carriers flowing from source to drain results in a positive current from drain to source, I_{DS}.

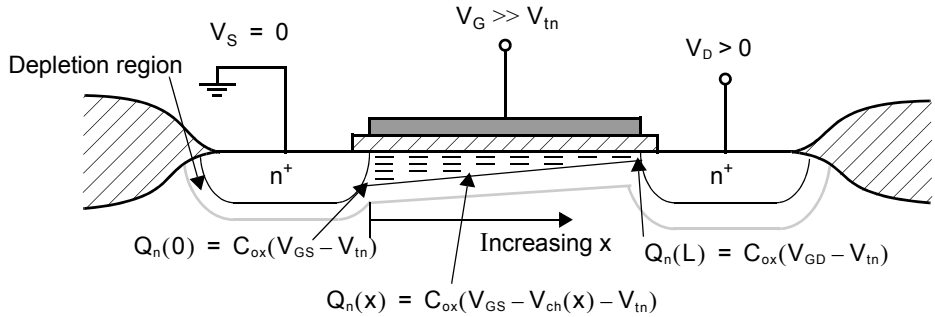


Fig. 1.11 The channel charge density for $V_{DS} > 0$.

As the drain-source voltage increases, the channel charge concentration decreases at the drain end. This decrease is due to the smaller gate-to-channel voltage difference across the thin gate oxide as one moves closer to the drain. In other words, since the drain voltage is assumed to be at a higher voltage than the source, there is an increasing voltage gradient from the source to the drain, resulting in a smaller gate-to-channel voltage near the drain. Since the charge density at a distance x from the source end of the channel is proportional to $V_G - V_{ch}(x) - V_{tn}$, as $V_G - V_{ch}(x)$ decreases, the charge density also decreases.[8] This effect is illustrated in Fig. 1.11.

Key Point: *The relationship between drain-source voltage and drain current of a MOS device is approximately linear when* $V_{DS} << V_{eff}$.

Note that at the drain end of the channel, we have

$$V_G - V_{ch}(L) = V_{GD} \tag{1.57}$$

For small V_{DS}, we saw from (1.56) that I_D was linearly related to V_{DS}. However, as V_{DS} increases, and the charge density decreases near the drain, the relationship becomes nonlinear. In fact, the linear relationship for I_D versus V_{DS} flattens for larger V_{DS}, as shown in Fig. 1.12.

As the drain voltage is increased, at some point the gate-to-channel voltage at the drain end will decrease to the threshold value V_{tn}—the minimum gate-to-channel voltage needed for n carriers in the channel to exist. Thus, at the drain end, the channel becomes *pinched off,* as shown in Fig. 1.13. This pinch-off occurs at $V_{GD} = V_{tn}$, since the channel voltage at the drain end is simply equal to V_D. Thus, pinch-off occurs for

$$V_{DG} > -V_{tn} \tag{1.58}$$

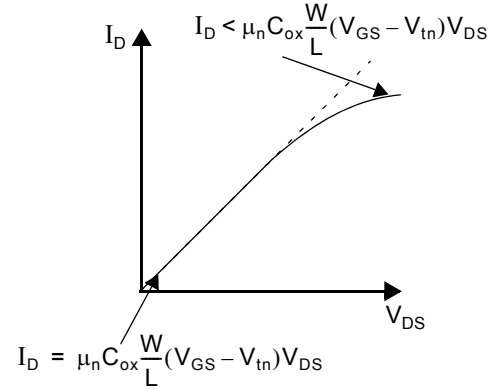


Fig. 1.12 For V_{DS} not close to zero, the I_D versus V_{DS} relationship is no longer linear.

Denoting $V_{DS\text{-}sat}$ as the drain-source voltage when the channel becomes pinched off, we can substitute $V_{DG} = V_{DS} - V_{GS}$ into (1.58) and find an equivalent pinch-off expression

$$V_{DS} > V_{DS\text{-}sat} \tag{1.59}$$

8. $V_G - V_{CH}(x)$ is the gate-to-channel voltage drop at distance x from the source end, with V_G being the same everywhere in the gate, since the gate material is highly conductive.

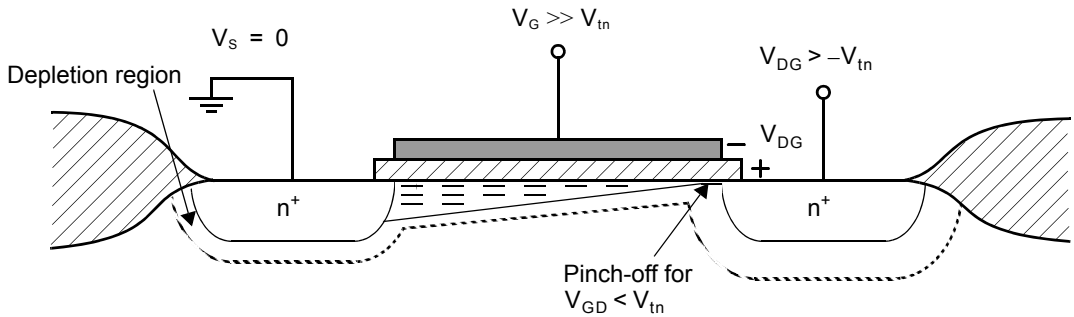


Fig. 1.13 When V_{DS} is increased so that $V_{GD} < V_{tn}$, the channel becomes pinched off at the drain end.

where $V_{DS\text{-}sat}$ is given[9] by

$$V_{DS\text{-}sat} = V_{GS} - V_{tn} = V_{eff} \tag{1.60}$$

The current travelling through the pinched-off channel region is saturated, similar to a gas under pressure travelling through a very small tube. If the drain-gate voltage rises above this critical pinch-off voltage of $-V_{tn}$, the charge concentration in the channel remains constant (to a first-order approximation) and the drain current no longer increases with increasing V_{DS}. The result is the current-voltage relationship shown in Fig. 1.14 for a given gate-source voltage. In the region of operation where $V_{DS} > V_{DS\text{-}sat}$, the drain current is independent of V_{DS} and is called the *active region*.[10] The region where I_D changes linearly with V_{DS} is called the *triode region*. When MOS transistors are used in analog amplifiers, they almost always are biased in the active region. When they are used in digital logic gates, they often operate in both regions.

It is sometimes necessary to distinguish between transistors in *weak, moderate,* and *strong inversion.* As just discussed, a gate-source voltage greater than V_{tn} results in an inverted channel, and drain-source current can flow. However, as the gate-source voltage is increased, the channel does not become inverted (i.e., n-region) suddenly,

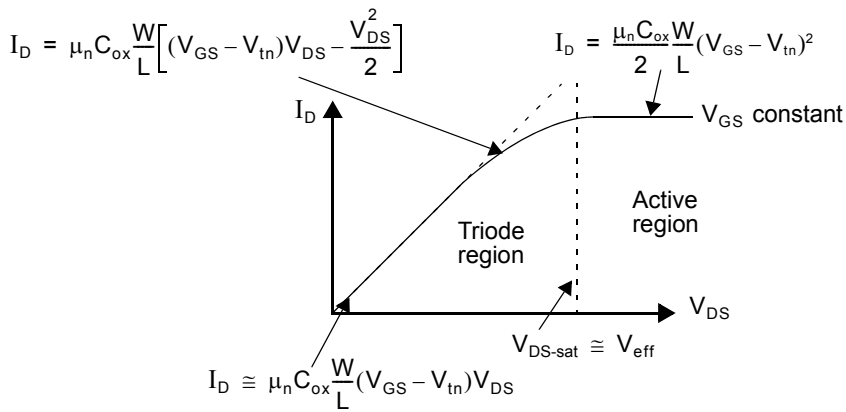


Fig. 1.14 The I_D versus V_{DS} curve for an ideal MOS transistor. For $V_{DS} > V_{DS\text{-}sat}$, I_D is approximately constant.

9. Because of the body effect, the threshold voltage at the drain end of the transistor is increased, resulting in the true value of $V_{DS\text{-}sat}$ being slightly lower than V_{eff}.

10. The active region may also be called the saturation region, but this can lead to confusion because in the case of bipolar transistors, the saturation region occurs for small V_{CE}, whereas for MOS transistors it occurs for large V_{DS}. Moreover, we shall see that the drain current does not truly saturate, but continues to increase slightly with increasing drain-source voltage.

but rather gradually. Thus, it is useful to define three regions of channel inversion with respect to the gate-source voltage. In most circuit applications, noncutoff MOSFET transistors are operated in strong inversion, with $V_{eff} > 100$ mV (many prudent circuit designers use a minimum value of 200 mV). As the name suggests, strong inversion occurs when the channel is strongly inverted. It should be noted that all the equation models in this section assume strong inversion operation. Weak inversion occurs when V_{GS} is approximately 100 mV or more below V_{tn} and is discussed as subthreshold operation in Section 1.4.1. Finally, moderate inversion is the region between weak and strong inversion.

1.2.3 Large-Signal Modelling

The *triode region equation* for a MOS transistor relates the drain current to the gate-source and drain-source voltages. It can be shown (see Appendix) that this relationship is given by

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{1.61}$$

As V_{DS} increases, I_D increases until the drain end of the channel becomes pinched off, and then I_D no longer increases. This pinch-off occurs for $V_{DG} = -V_{tn}$, or approximately,

$$V_{DS} = V_{GS} - V_{tn} = V_{eff} \tag{1.62}$$

Right at the edge of pinch-off, the drain current resulting from (1.61) and the drain current in the active region (which, to a first-order approximation, is constant with respect to V_{DS}) must have the same value. Therefore, the *active region equation* can be found by substituting (1.62) into (1.61), resulting in

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2 \tag{1.63}$$

For $V_{DS} > V_{eff}$, the current stays constant at the value given by (1.63), ignoring second-order effects such as channel-length modulation. This equation is perhaps the most important one that describes the large-signal operation of a MOS transistor. It should be noted here that (1.63) represents a square-law current–voltage relationship for a MOS transistor in the active region. In the case of a BJT transistor, an exponential current–voltage relationship exists in the active region.

> ***Key Point:*** *For $V_{DS} > V_{eff}$, a MOS device exhibits a square-law current–voltage relationship.*

As just mentioned, (1.63) implies that the drain current, I_D, is independent of the drain-source voltage. This independence is only true to a first-order approximation. The major source of error is due to the channel length shrinking as V_{DS} increases. To see this effect, consider Fig. 1.15, which shows a cross section of a transistor in the

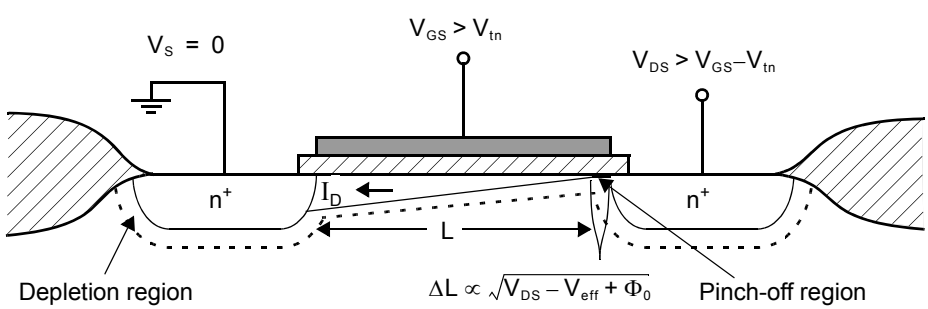


Fig. 1.15 Channel length shortening for $V_{DS} > V_{eff}$.

active region. A pinched-off region with very little charge exists between the drain and the channel. The voltage at the end of the channel closest to the drain is fixed at $V_{GS} - V_{tn} = V_{eff}$. The voltage difference between the drain and the near end of the channel lies across a short depletion region often called the *pinch-off region*. As V_{DS} becomes larger than V_{eff}, this depletion region surrounding the drain junction increases its width in a square-root relationship with respect to V_{DS}. This increase in the width of the depletion region surrounding the drain junction decreases the effective channel length. In turn, this decrease in effective channel length increases the drain current, resulting in what is commonly referred to as *channel-length modulation*.

To derive an equation to account for channel-length modulation, we first make use of (1.11) and denote the width of the depletion region by x_d, resulting in

$$\begin{aligned} x_d &\cong k_{ds}\sqrt{V_{D\text{-}ch} + \Phi_0} \\ &= k_{ds}\sqrt{V_{DG} + V_{tn} + \Phi_0} \end{aligned} \tag{1.64}$$

where

$$k_{ds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}} \tag{1.65}$$

and has units of $m/\sqrt{V}$. Note that N_A is used here since the n-type drain region is more heavily doped than the p-type channel (i.e., $N_D >> N_A$). By writing a Taylor approximation for I_D around its operating value of $V_{DS} = V_{GS} - V_{tn} = V_{eff}$, we find I_D to be given by

$$I_D = I_{D\text{-}sat} + \left(\frac{\partial I_D}{\partial L}\right)\left(\frac{\partial L}{\partial V_{DS}}\right)\Delta V_{DS} \cong I_{D\text{-}sat}\left(1 + \frac{k_{ds}(V_{DS} - V_{eff})}{2L\sqrt{V_{DG} + V_{tn} + \Phi_0}}\right) \tag{1.66}$$

where $I_{D\text{-}sat}$ is the drain current when $V_{DS} = V_{eff}$, or equivalently, the drain current when the channel-length modulation is ignored. Note that in deriving the final equation of (1.66), we have used the relationship $\partial L/\partial V_{DS} = -\partial x_d/\partial V_{DS}$. Usually, (1.66) is written as

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2[1 + \lambda(V_{DS} - V_{eff})] \tag{1.67}$$

where λ is the output impedance constant (in units of V^{-1}) given by

$$\lambda = \frac{k_{ds}}{2L\sqrt{V_{DG} + V_{tn} + \Phi_0}} = \frac{k_{ds}}{2L\sqrt{V_{DS} - V_{eff} + \Phi_0}} \tag{1.68}$$

Equation (1.67) is accurate until V_{DS} is large enough to cause second-order effects, often called *short-channel effects*. For example, (1.67) assumes that current flow down the channel is not *velocity-saturated*, in which case increasing the electric field no longer increases the carrier speed. Short-channel effects cause I_D to deviate from the result predicted by (1.67) and are discussed in Section 1.4. Of course, for quite large values of V_{DS}, the transistor will eventually break down.

A plot of I_D versus V_{DS} for different values of V_{GS} is shown in Fig. 1.16. Note that in the active region, the small (but nonzero) slope indicates the small dependence of I_D on V_{DS}.

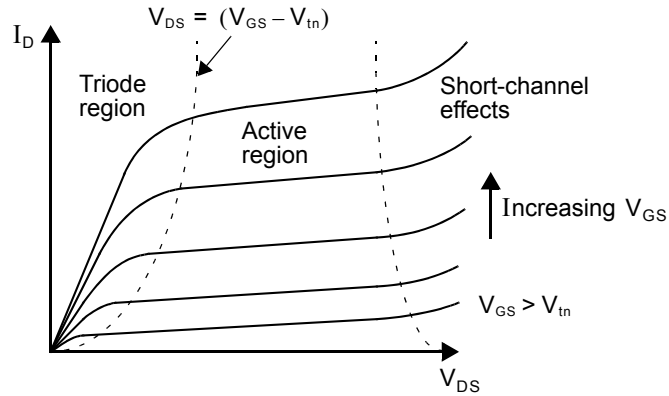


Fig. 1.16 I_D versus V_{DS} for different values of V_{GS}.

EXAMPLE 1.8

Find I_D for an n-channel transistor that has doping concentrations of $N_D = 10^{25}$ electrons/m^3, $N_A = 5 \times 10^{22}$ holes/m^3, $\mu_n C_{ox} = 270\ \mu A/V^2$, $W/L = 5\ \mu m/0.5\ \mu m$, $V_{GS} = 0.8$ V, $V_{tn} = 0.45$ V, and $V_{DS} = V_{eff}$. Assuming λ remains constant, estimate the new value of I_D if V_{DS} is increased by 0.5 V.

Solution

From (1.65), we have

$$k_{ds} = \sqrt{\frac{2 \times 11.8 \times 8.854 \times 10^{-12}}{1.6 \times 10^{-19} \times 5 \times 10^{22}}} = 162 \times 10^{-9}\ \text{m}/\sqrt{\text{V}}$$

which is used in (1.68) to find λ as

$$\lambda = \frac{162 \times 10^{-9}}{2 \times 0.5 \times 10^{-6} \times \sqrt{0.9}} = 0.171\ \text{V}^{-1}$$

Using (1.67), we find for $V_{DS} = V_{eff} = 0.4$ V,

$$I_{D1} = \left(\frac{270 \times 10^{-6}}{2}\right)\left(\frac{5}{0.5}\right)(0.35)^2(1) = 165\ \mu\text{A}$$

In the case where $V_{DS} = V_{eff} + 0.5$ V $= 0.9$ V, we have

$$I_{D2} = 165\ \mu\text{A} \times (1 + \lambda \times 0.5) = 180\ \mu\text{A}$$

Note that this example shows almost a 10 percent increase in drain current for a 0.5 V increase in drain-source voltage.

1.2.4 Body Effect

Key Point: *The body effect is the influence of the body potential on the channel, modelled as an increase in the threshold voltage, V_{tn}, with increasing source-to-body reverse-bias.*

The large-signal equations in the preceding section were based on the assumption that the source voltage was the same as the body voltage (i.e., the substrate or bulk voltage or an NMOS device). However, often the source and body can be at different voltages. Looking at Fig. 1.11, it is evident that the body is capacitively coupled to the channel region just as the gate is, albeit through the junction capacitance between them instead of the gate-oxide capacitance.[11] Hence, the amount of charge in the channel and conduction through it is influenced by the potential difference between body and source.

Typically called the *body effect*, the influence of the body potential on the channel is modelled as an increase in the threshold voltage, V_{tn}, with increasing source-to-body reverse-bias voltage. The body effect is more important for transistors in a well of a CMOS process where the body terminal's doping concentration is higher and the resulting junction capacitance between body and channel is larger. The body effect is often important in analog circuit designs and should not be ignored.

To account for the body effect, it can be shown (see Appendix at the end of this chapter) that the threshold voltage of an n-channel transistor is now given by

$$V_{tn} = V_{tn0} + \gamma\left(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right) \tag{1.69}$$

where V_{tn0} is the threshold voltage with zero V_{SB} (source-to-body voltage), $\phi_F = (kT/q)\ln(N_A/n_i)$ is the Fermi potential of the body, and

$$\gamma = \frac{\sqrt{2qN_AK_s\varepsilon_0}}{C_{ox}} \tag{1.70}$$

The factor γ is often called the *body-effect constant* and has units of $\sqrt{V}$. Notice that γ is proportional to $\sqrt{N_A}$,[12] so the body effect is larger for transistors in a well where typically the doping is higher than the substrate of the microcircuit.

1.2.5 p-Channel Transistors

All of the preceding equations have been presented for n-channel enhancement transistors. *In the case of p-channel transistors, these equations can also be used if a negative sign is placed in front of every voltage variable.* Thus, V_{GS} becomes V_{SG}, V_{DS} becomes V_{SD}, V_{tn} becomes $-V_{tp}$, and so on. The condition required for conduction is now $V_{SG} > V_{tp}$, where V_{tp} is now a negative quantity for an enhancement p-channel transistor.[13] The requirement on the source-drain voltage for a p-channel transistor to be in the active region is $V_{SD} > V_{SG} + V_{tp}$. The equations for I_D, in both regions, remain unchanged, because all voltage variables are squared, resulting in positive hole current flow from the source to the drain in p-channel transistors.

---

11. In fact, JFETs intentionally modulate the conducting channel via a junction capacitance, hence their name: Junction Field-Effect Transistors.
12. For an n-channel transistor. For a p-channel transistor, γ is proportional to the square root of N_D.
13. It is possible to realize depletion p-channel transistors, but these are of little value and seldom worth the extra processing involved. Depletion n-channel transistors are also seldom encountered in CMOS microcircuits, although they might be worth the extra processing involved in some applications, especially if they were in a well.

1.2.6 Low-Frequency Small-Signal Modelling in the Active Region

The most commonly used low-frequency small-signal model for a MOS transistor operating in the active region shown in Fig. 1.17. The voltage-controlled current source, $g_m v_{gs}$, is the most important component of the model, with the transistor transconductance g_m defined as

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{1.71}$$

In the active region, we use (1.63), which is repeated here for convenience,

$$I_D = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2 = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}^{\ 2} \tag{1.72}$$

and we apply the derivative shown in (1.71) to obtain

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{tn}) = \mu_n C_{ox}\frac{W}{L}V_{eff} \tag{1.73}$$

or equivalently,

$$g_m = \mu_n C_{ox}\frac{W}{L}V_{eff} \tag{1.74}$$

where the effective gate-source voltage, V_{eff}, is defined as $V_{eff} \equiv V_{GS} - V_{tn}$. Thus, we see that the transconductance of a MOS transistor is directly proportional to V_{eff}.

Sometimes it is desirable to express g_m in terms of I_D rather than V_{GS}. From (1.72), we have

$$V_{GS} = V_{tn} + \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{1.75}$$

The second term in (1.75) is the effective gate-source voltage, V_{eff}, where

$$V_{eff} = V_{GS} - V_{tn} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{1.76}$$

Substituting (1.76) in (1.74) results in an alternative expression for g_m.

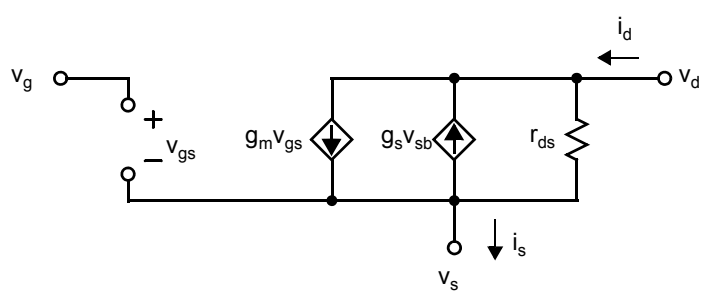


Fig. 1.17 The low-frequency, small-signal model for an active MOS transistor.

$$g_m = \sqrt{2\mu_n C_{ox}\frac{W}{L}I_D} \tag{1.77}$$

Thus, the transistor transconductance is proportional to $\sqrt{I_D}$ for a MOS transistor.[14]

A third expression for g_m is found by rearranging (1.77) and then using (1.76) to obtain

$$g_m = \frac{2I_D}{V_{eff}} \tag{1.78}$$

> ***Key Point:*** *The square-root relationship for transconductance* $g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$ is useful for circuit analysis when device sizes are fixed. However, the simpler expression $g_m = 2I_D/V_{eff}$ is useful during initial circuit design when transistor sizes are yet to be determined.

Note that this expression is independent of $\mu_n C_{ox}$ and W/L, and it relates the transconductance to the ratio of drain current to effective gate-source voltage. However, it can cause some confusion for new analog designers as it appears to indicate that transconductance is proportional to drain current, whereas (1.77) indicates a square-root relationship. This discrepancy is resolved by recognizing that increasing I_D while keeping V_{eff} constant implies a proportional increase in (W/L). Hence, (1.77) is useful for analysis where the transistor sizes are given whereas the simple expression in (1.78) can be quite useful during an initial circuit design when the transistor sizes are yet to be determined.

Design often begins with a set of transistor voltage–current measurements, obtained either by experiment or simulation, from which the designer may estimate the device constants $\mu_n C_{ox}$, V_{tn}, etc. based on the relationships presented above.

EXAMPLE 1.9

The drain current for a particular NMOS device with an aspect ratio $W/L = 10$ is plotted in Fig. 1.18(*a*) versus V_{GS} for constant drain, source, and body voltages. From this data, estimate $\mu_n C_{ox}$ and V_{tn}.

Solution

Taking the derivative of Fig. 1.18(*a*) gives the plot of $g_m = \delta I_D/\delta V_{GS}$ in Fig. 1.18(*b*). Based on the square-law equation (1.74), this plot should be linear in the active region intersecting the line $g_m = 0$ at $V_{eff} = 0$. Hence, extrapolating the linear portion of the curve in Fig. 1.18(*b*) provides an intersection at $V_{GS} = V_{tn}$. In this case, $V_{tn} \cong 450$ mV. Furthermore, (1.73) shows that the slope of the g_m versus V_{GS} curve should be $\mu_n C_{ox}(W/L)$ in active operation. In Fig. 1.18(*b*) this slope is approximately 2.7 mA/V^2 which translates to a value of $\mu_n C_{ox} = 270\ \mu$A/V^2. These are approximate values, particularly since (1.74) is derived without considering channel-length modulation. However, these values are very useful during design for quickly and roughly estimating the device sizes, currents and voltages required to obtain a desired g_m.

The second voltage-controlled current-source in Fig. 1.17, shown as $g_s v_s$, models the body effect on the small-signal drain current, i_d. When the source is connected to small-signal ground, or when its voltage does not change appreciably, then this current source can be ignored. When the body effect cannot be ignored, we have

$$g_s = \frac{\partial I_D}{\partial V_{SB}} = \frac{\partial I_D}{\partial V_{tn}}\frac{\partial V_{tn}}{\partial V_{SB}} \tag{1.79}$$

14. Whereas it is proportional to I_C for a BJT.

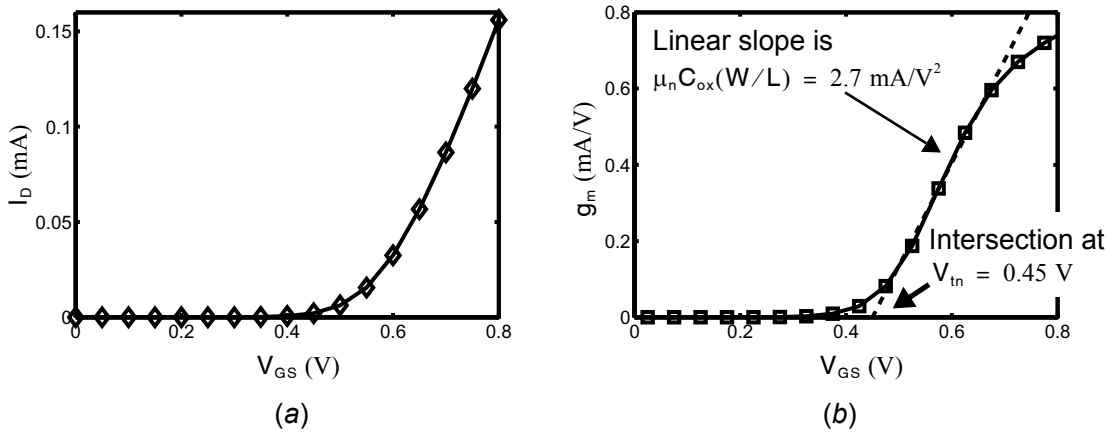


Fig. 1.18 Estimation of V_{tn} and $\mu_n C_{ox}$ from transistor voltage–current data.

From (1.72) we have

$$\frac{\partial I_D}{\partial V_{tn}} = -\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn}) = -g_m \tag{1.80}$$

Using (1.69), which gives V_{tn} as

$$V_{tn} = V_{tn0} + \gamma(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}) \tag{1.81}$$

we have

$$\frac{\partial V_{tn}}{\partial V_{SB}} = \frac{\gamma}{2\sqrt{V_{SB} + |2\phi_F|}} \tag{1.82}$$

The negative sign of (1.80) is eliminated by subtracting the current $g_s v_s$ from the major component of the drain current, $g_m v_{gs}$, as shown in Fig. 1.17. Thus, using (1.80) and (1.82), we have

$$g_s = \frac{\gamma g_m}{2\sqrt{V_{SB} + |2\phi_F|}} \tag{1.83}$$

Note that although g_s is nonzero for $V_{SB} = 0$, if the source is connected to the bulk v_{sb} is zero and g_s may be excluded from the model. However, if the source happens to be biased at the same potential as the bulk but is not directly connected to it, then the effect of g_s should be taken into account since there may be nonzero small signals, v_{sb}.

The resistor, r_{ds}, shown in Fig. 1.17, accounts for the finite output impedance (i.e., it models the channel-length modulation and its effect on the drain current due to changes in V_{DS}). Using (1.67), repeated here for convenience,

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2[1 + \lambda(V_{DS} - V_{eff})] \tag{1.84}$$

we have

$$\frac{1}{r_{ds}} = g_{ds} = \frac{\partial I_D}{\partial V_{DS}} = \lambda\left(\frac{\mu_n C_{ox}}{2}\right)\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2 = \lambda I_{D\text{-}sat} \cong \lambda I_D \tag{1.85}$$

where the approximation assumes λ is small, such that we can approximate the drain bias current as being the same as $I_{D\text{-}sat}$. Thus,

$$r_{ds} \cong \frac{1}{\lambda I_D} \tag{1.86}$$

where

$$\lambda = \frac{k_{ds}}{2L\sqrt{V_{DS} - V_{eff} + \Phi_0}} \tag{1.87}$$

and

$$k_{ds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}} \tag{1.88}$$

Key Point: *Small signal* r_{ds} *is proportional to* L/I_D.

Substituting (1.87) into (1.86) reveals that r_{ds} is proportional to L/I_D. It should be noted here that (1.86) is often empirically adjusted to take into account second-order effects.

EXAMPLE 1.10

Derive the low-frequency model parameters for an n-channel transistor that has doping concentrations of $N_D = 10^{25}$ electrons/m^3, $N_A = 5 \times 10^{22}$ holes/m^3, $\mu_n C_{ox} = 270\ \mu A/V^2$, $W/L = 5\ \mu m/0.5\ \mu m$, $V_{GS} = 0.8$ V, $V_{tn} = 0.45$ V, and $V_{DS} = V_{eff}$. Assume $\gamma = 0.25\ \sqrt{V}$ and $V_{SB} = 0.5$ V. What is the new value of r_{ds} if the drain-source voltage is increased by 0.5 V?

Solution

Since these parameters are the same as in Example 1.8, we have

$$g_m = \frac{2I_D}{V_{eff}} = \frac{2 \times 165\ \mu A}{0.35\ V} = 0.94\ mA/V$$

The Fermi potential of the body at room temperature with $N_A = 5 \times 10^{22}$ holes/m^3 is $\phi_F = (kT/q)\ln(N_A/n_i) \cong 0.38$ V, and from (1.83) we have

$$g_s = \frac{0.25 \times 0.94 \times 10^{-3}}{2\sqrt{0.5 + 0.766}} = 0.104\ mA/V$$

Note that this source-bulk transconductance value is about 1/9th that of the gate-source transconductance. For r_{ds}, we use (1.86) to find

$$r_{ds} = \frac{1}{0.171 \times 165 \times 10^{-6}} = 35\ \text{k}\Omega$$

Recalling that $V_{eff} = 0.35$ V, if V_{DS} is increased to 0.85 V, the new value for λ is

$$\lambda = \frac{162 \times 10^{-9}}{2(0.5 \times 10^{-6})\sqrt{1.4}} = 0.137\ \text{V}^{-1}$$

resulting in a new value of r_{ds} given by

$$r_{ds} = \frac{1}{\lambda I_{D2}} = \frac{1}{0.137 \times 180 \times 10^{-6}} \cong 41\ \text{k}\Omega$$

EXAMPLE 1.11

Plotted in Fig. 1.19(*a*) is the drain current of a particular NMOS device versus its V_{DS} with constant gate, source, and body voltages. From this data, estimate the device parameter λ.

Solution

First, rearrange (1.85) into

$$\lambda = \frac{g_{ds}}{I_{D\text{-}sat}} = \left(\frac{\partial I_D}{\partial V_{DS}}\right)\frac{1}{I_{D\text{-}sat}}.$$

Hence, a plot of $g_{ds}/I_D = (\partial I_D/\partial V_{DS})/I_D$ versus V_{DS} for a constant value of V_{GS} will be roughly flat in strong inversion providing a rough estimate of λ. This is done for the present example in Fig. 1.19(*b*) yielding a value of $\lambda = 0.45\ \text{V}^{-1}$. Estimates obtained in this way are only approximate since r_{ds} is subject to many higher-order effects, especially at short channel lengths where λ may change by 50% or even more.

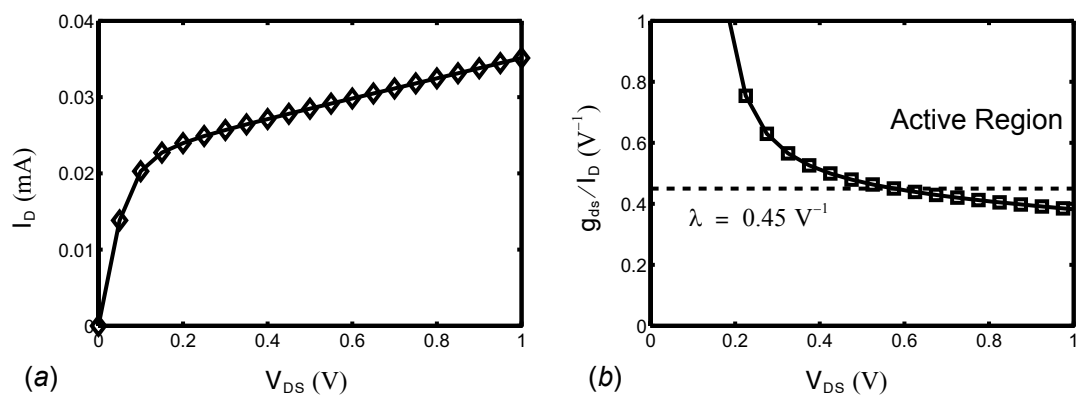


Fig. 1.19 Estimation of λ from transistor voltage–current data.

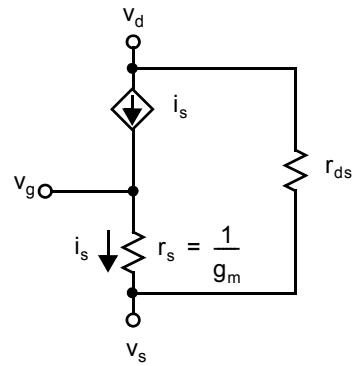


Fig. 1.20 The small-signal, low-frequency T model for an active MOS transistor (the body effect is not modelled).

An alternate low-frequency model, known as a T model, is shown in Fig. 1.20. This *T model* can often result in simpler equations and is most often used by experienced designers for a quick analysis. At first glance, it might appear that this model allows for nonzero gate current, but a quick check confirms that the drain current must always equal the source current, and, therefore, the gate current must always be zero. For this reason, when using the T model, one assumes from the beginning that the gate current is zero.

EXAMPLE 1.12

Find the T model parameter, r_s, for the transistor in Example 1.10.

Solution

The value of r_s is simply the inverse of g_m, resulting in

$$r_s = \frac{1}{g_m} = \frac{1}{0.94 \times 10^{-3}} = 1.06\ \text{k}\Omega$$

The value of r_{ds} remains the same, either 35 kΩ or 41 kΩ, depending on the drain-source voltage.

1.2.7 High-Frequency Small-Signal Modelling in the Active Region

A high-frequency model of a MOSFET in the active region is shown in Fig. 1.21. Most of the capacitors in the small-signal model are related to the physical transistor. Shown in Fig. 1.22 is a cross section of a MOS transistor, where the parasitic capacitances are shown at the appropriate locations. The largest capacitor in Fig. 1.22 is C_{gs}. This capacitance is primarily due to the change in channel charge as a result of a change in V_{GS}. It can be shown [Tsividis, 1987] that C_{gs} is approximately given by

$$C_{gs} \cong \frac{2}{3} WLC_{ox} \tag{1.89}$$

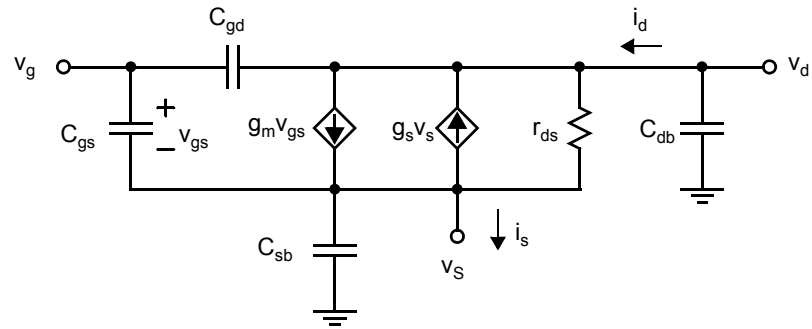


Fig. 1.21 The small-signal model for a MOS transistor in the active region.

When accuracy is important, an additional term should be added to (1.89) to take into account the overlap between the gate and source junction, which should include the *fringing capacitance* (fringing capacitance is due to boundary effects). This additional component is given by

$$C_{ov} = WL_{ov}C_{ox} \tag{1.90}$$

where L_{ov} is the effective overlap distance and is usually empirically derived[15]. Thus,

$$C_{gs} = \frac{2}{3}WLC_{ox} + C_{ov} \tag{1.91}$$

when higher accuracy is needed.

Key Point: *In a MOSFET, the largest parasitic capacitance is* C_{gs}*, proportional to gate area* WL *and via* C_{ox} *inversely proportional to oxide thickness.*

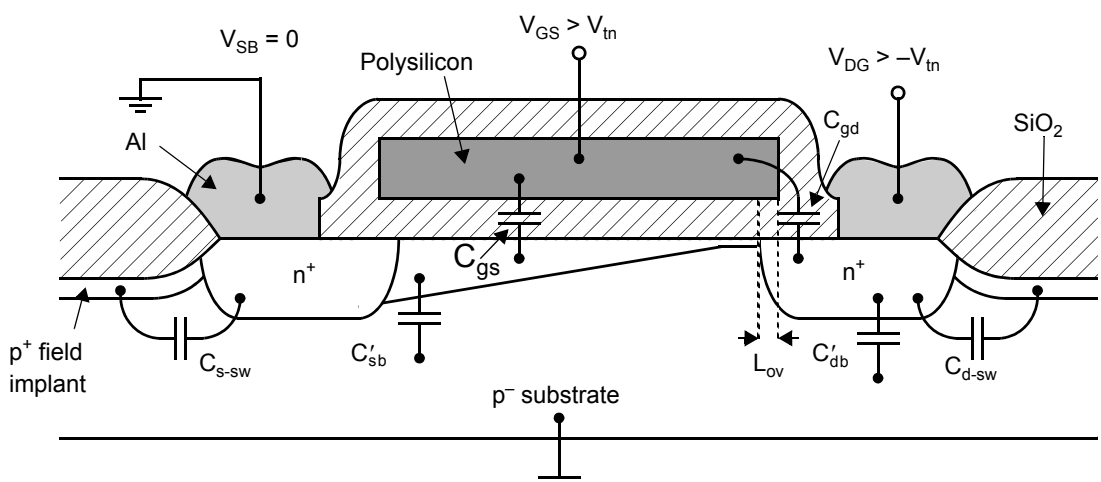


Fig. 1.22 A cross section of an n-channel MOS transistor showing the small-signal capacitances.

15. Part of the overlap capacitance is due to *fringe* electric fields, and therefore L_{ov} is usually taken larger than its actual physical overlap to more accurately give an *effective* value for overlap capacitances.

The next largest capacitor in Fig. 1.22 is C'_{sb}, the capacitor between the source and the substrate. This capacitor is due to the depletion capacitance of the reverse-biased source junction, and it includes the channel-to-bulk capacitance (assuming the transistor is on). Its size is given by

$$C'_{sb} = (A_s + A_{ch})C_{js} \tag{1.92}$$

where A_s is the area of the source junction, A_{ch} is the area of the channel (i.e., WL) and C_{js} is the depletion capacitance of the source junction, given by

$$C_{js} = \frac{C_{j0}}{\sqrt{1 + \dfrac{V_{SB}}{\Phi_0}}} \tag{1.93}$$

Note that the total area of the effective source includes the original area of the junction (when no channel is present) plus the effective area of the channel.

The depletion capacitance of the drain is smaller because it does not include the channel area. Here, we have

$$C'_{db} = A_d C_{jd} \tag{1.94}$$

where

$$C_{jd} = \frac{C_{j0}}{\sqrt{1 + \dfrac{V_{DB}}{\Phi_0}}} \tag{1.95}$$

and A_d is the area of the drain junction.

The capacitance C_{gd}, sometimes called the Miller capacitance, is important when there is a large voltage gain between gate and drain. It is primarily due to the overlap between the gate and the drain and fringing capacitance. Its value is given by

$$C_{gd} = C_{ox}WL_{ov} \tag{1.96}$$

Key Point: *The gate-drain capacitance C_{gd}, also known as the Miller capacitance, is due to physical overlap of the gate and drain regions as well as fringing fields. It is especially important when there is a large voltage gain between gate and drain.*

where, once again, L_{ov} is usually empirically derived.

Two other capacitors are often important in integrated circuits. These are the source and drain *sidewall capacitances,* $C_{s\text{-}sw}$ and $C_{d\text{-}sw}$. These capacitances can be large because of some highly doped p^+ regions under the thick field oxide called *field implants.* The major reason these regions exist is to ensure there is no leakage current between transistors. Because they are highly doped and they lie beside the highly doped source and drain junctions, the sidewall capacitances can result in large additional capacitances that must be taken into account in determining C_{sb} and C_{db}. The sidewall capacitances are especially important in modern technologies as dimensions shrink. For the source, the sidewall capacitance is given by

$$C_{s\text{-}sw} = P_s C_{j\text{-}sw} \tag{1.97}$$

where P_s is the length of the perimeter of the source junction, excluding the side adjacent to the channel, and

$$C_{j\text{-}sw} = \frac{C_{j\text{-}sw0}}{\sqrt{1 + \dfrac{V_{SB}}{\Phi_0}}} \tag{1.98}$$

It should be noted that $C_{j\text{-}sw0}$, the sidewall capacitance per unit length at 0-V bias voltage, can be quite large because the field implants are heavily doped.

The situation is similar for the drain sidewall capacitance, $C_{d\text{-}sw}$,

$$C_{d\text{-}sw} = P_d C_{j\text{-}sw} \tag{1.99}$$

where P_d is the drain perimeter excluding the portion adjacent to the gate.

Finally, the source-bulk capacitance, C_{sb}, is given by

$$C_{sb} = C'_{sb} + C_{s\text{-}sw} \tag{1.100}$$

with the drain-bulk capacitance, C_{db}, given by

$$C_{db} = C'_{db} + C_{d\text{-}sw} \tag{1.101}$$

EXAMPLE 1.13

An n-channel transistor is modelled as having the following capacitance parameters: $C_j = 2.4 \times 10^{-4}$ pF/(μm)^2, $C_{j\text{-}sw} = 2.0 \times 10^{-4}$ pF/μm, $C_{ox} = 1.9 \times 10^{-3}$pF/(μm)^2, $C_{ov} = 2.0 \times 10^{-4}$ pF/μm. Find the capacitances C_{gs}, C_{gd}, C_{db}, and C_{sb} for a transistor having $W = 100$ μm and $L = 2$ μm. Assume the source and drain junctions extend 4 μm beyond the gate, so that the source and drain areas are $A_s = A_d = 400$ (μm)^2 and the perimeter of each is $P_s = P_d = 108$ μm.

Solution

We calculate the various capacitances as follows:

$$C_{gs} = \left(\frac{2}{3}\right) WLC_{ox} + C_{ov} \times W = 0.27 \text{ pF}$$

$$C_{gd} = C_{ov} \times W = 0.02 \text{ pF}$$

$$C_{sb} = C_j(A_s + WL) + (C_{j\text{-}sw} \times P_s) = 0.17 \text{ pF}$$

$$C_{db} = (C_j \times A_d) + (C_{j\text{-}sw} \times P_d) = 0.12 \text{ pF}$$

Note that the source-bulk and drain-bulk capacitances are significant compared to the gate-source capacitance, in this case 1 – 2 fF/μm width compared with 2.7 fF/μm for C_{gs}. Thus, for high-speed circuits, it is important to keep the areas and perimeters of drain and source junctions as small as possible (possibly by sharing junctions between transistors, as seen in the next chapter).

1.2.8 Small-Signal Modelling in the Triode and Cutoff Regions

The low-frequency, small-signal model of a MOS transistor in the triode region (which is sometimes referred to as the linear region) is a voltage-controlled resistor with V_{GS}, or equivalently V_{eff}, used as the control terminal. Using (1.61), the large-signal equation for I_D in the triode region,

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{1.102}$$

results in

$$\frac{1}{r_{ds}} = g_{ds} = \frac{\partial I_D}{\partial V_{DS}} = \mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn} - V_{DS}) \tag{1.103}$$

where r_{ds} is the small-signal drain-source resistance (and g_{ds} is the conductance). For the common case of V_{DS} near zero, we have

$$g_{ds} = \frac{1}{r_{ds}} = \mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn}) = \mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff} \tag{1.104}$$

which is similar to the I_D-versus-V_{DS} relationship given earlier in (1.56).

EXAMPLE 1.14

For the transistor of Example 1.10, find the triode model parameters when V_{DS} is near zero.

Solution

From (1.104), we have

$$g_{ds} = 270 \times 10^{-6} \times \left(\frac{5}{0.5}\right) \times 0.35 = 0.94 \text{ mA/V}$$

Note that this conductance value is the same as the transconductance of the transistor, g_m, in the active region. The resistance, r_{ds}, is simply $1/g_{ds}$, resulting in $r_{ds} = 1.06$ kΩ.

The accurate small-signal modelling of the high-frequency operation of a transistor in the triode region is nontrivial (even with the use of a computer simulation). A moderately accurate model is shown in Fig. 1.23, where the gate-to-channel capacitance and the channel-to-substrate capacitance are modelled as distributed elements. However, the I-V relationships of the distributed RC elements are highly nonlinear because the junction capacitances of the source and drain are nonlinear depletion capacitances, as is the channel-to-substrate capacitance. Also, if V_{DS} is not small, then the channel resistance per unit length should increase as one moves closer to the drain. This model is much too complicated for use in hand analysis.

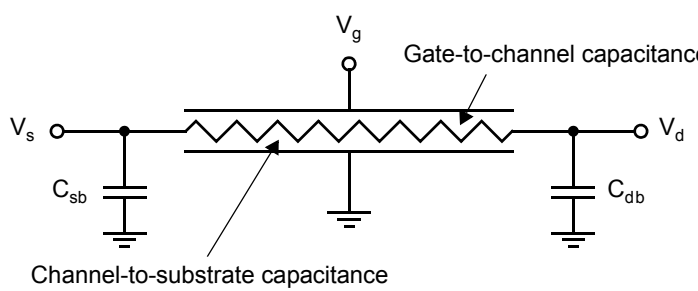


Fig. 1.23 A distributed RC model for a transistor in the triode region.

A simplified model often used for small V_{DS} is shown in Fig. 1.24, where the resistance, r_{ds}, is given by (1.104). Here, the gate-to-channel capacitance has been evenly divided between the source and drain nodes,

$$C_{gs} = C_{gd} = \frac{A_{ch}C_{ox}}{2} = \frac{WLC_{ox}}{2} \tag{1.105}$$

Note that this equation ignores the gate-to-junction overlap capacitances, as given by (1.90), which should be taken into account when accuracy is very important. The channel-to-substrate capacitance has also been divided in half and shared between the source and drain junctions. Each of these capacitors should be added to the junction-to-substrate capacitance and the junction-sidewall capacitance at the appropriate node. Thus, we have

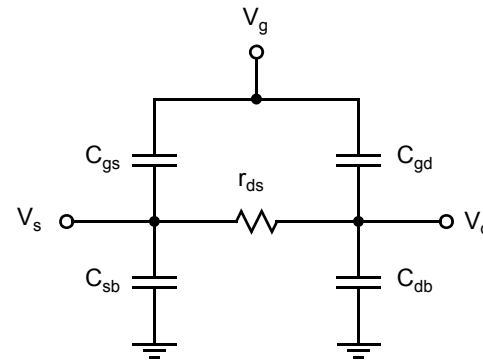


Fig. 1.24 A simplified triode-region model valid for small V_{DS}.

$$C_{sb\text{-}0} = C_{j0}\left(A_s + \frac{A_{ch}}{2}\right) + C_{j\text{-}sw0}P_s \tag{1.106}$$

and

$$C_{db\text{-}0} = C_{j0}\left(A_d + \frac{A_{ch}}{2}\right) + C_{j\text{-}sw0}P_d \tag{1.107}$$

Also,

$$C_{sb} = \frac{C_{sb\text{-}0}}{\sqrt{1 + \frac{V_{sb}}{\Phi_0}}} \tag{1.108}$$

and

$$C_{db} = \frac{C_{db\text{-}0}}{\sqrt{1 + \frac{V_{db}}{\Phi_0}}} \tag{1.109}$$

It might be noted that C_{sb} is often comparable in size to C_{gs} due to its larger area and the sidewall capacitance.

When the transistor turns off, the small-signal model changes considerably. A reasonable model is shown in Fig. 1.25. Perhaps the biggest difference is that r_{ds} is now infinite. Another major difference is that C_{gs} and C_{gd} are now much smaller. Since the channel has disappeared, these capacitors are now due to only overlap and fringing capacitance. Thus, we have

$$C_{gs} = C_{gd} = WL_{ov}C_{ox} \tag{1.110}$$

However, the reduction of C_{gs} and C_{gd} does not mean that the total gate capacitance is necessarily smaller. We now have a "new" capacitor, C_{gb}, which is the gate-to-substrate capacitance. This capacitor is highly nonlinear and dependent on the gate voltage. If the gate voltage has been very negative for some time and the gate is accumulated, then we have

$$C_{gb} = A_{ch}C_{ox} = WLC_{ox} \tag{1.111}$$

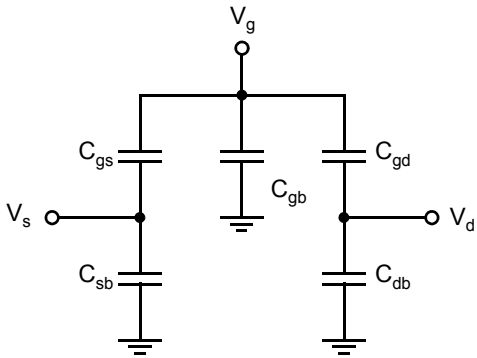


Fig. 1.25 A small-signal model for a MOSFET that is turned off.

If the gate-to-source voltage is around 0 V, then C_{gb} is equal to C_{ox} in series with the channel-to-bulk depletion capacitance and is considerably smaller, especially when the substrate is lightly doped. Another case where C_{gb} is small is just after a transistor has been turned off, before the channel has had time to accumulate. Because of the complicated nature of correctly modelling C_{gb} when the transistor is turned off, equation (1.111) is usually used for hand analysis as a worst-case estimate.

The capacitors C_{sb} and C_{db} are also smaller when the channel is not present. We now have

$$C_{sb\text{-}0} = A_s C_{j0} \tag{1.112}$$

and

$$C_{db\text{-}0} = A_d C_{j0} \tag{1.113}$$

1.2.9 Analog Figures of Merit and Trade-offs

With so many device constants and small-signal model parameters, it is sometimes useful to have a single number that captures some key aspect of transistor performance. Two such figures of merit are considered in this section: intrinsic gain, which relates to the transistor's low-frequency small-signal performance; and intrinsic speed, related to the transistor's high-frequency small-signal performance.

The *intrinsic gain* of a transistor is a measure of the maximum possible low-frequency small-signal voltage gain it can provide. The voltage gain of a transistor is generally maximized by operating it in the active mode with the source terminal (small-signal) grounded, the input applied to the gate, and the output observed at the drain terminal, as shown in Fig. 1.26(*a*). In order to maximize gain, an ideal current source is used to provide the drain current so that the only load the drain terminal sees is the transistor itself. Measured in this way, the intrinsic gain is closely related to the gain provided by several important single-stage amplifier stages discussed in later sections, such as common-source amplifiers and differential pairs with active loads.

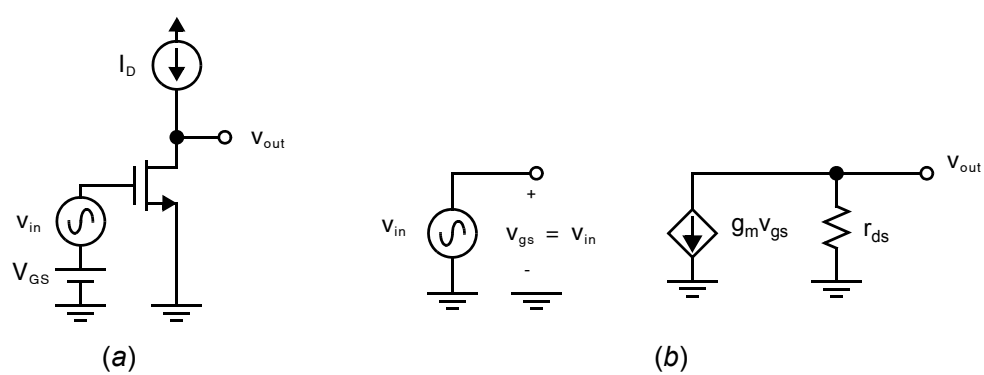


Fig. 1.26 The circuit used to characterize transistor intrinsic gain: *a*) complete circuit; *b*) dc small-signal equivalent.

The small-signal equivalent circuit at dc simplifies to that shown in Fig. 1.26(*b*). The transistor's intrinsic gain is therefore,

$$A_i = \left|\frac{v_{out}}{v_{in}}\right| = g_m r_{ds} \tag{1.114}$$

Substituting the expressions for g_m from equation (1.78) and r_{ds} from (1.86) gives a simple expression for transistor intrinsic gain dependent on only a few device parameters.

$$A_i \cong \frac{2I_D}{V_{eff}} \cdot \frac{1}{\lambda I_D} = \frac{2}{\lambda V_{eff}} \tag{1.115}$$

The intrinsic gain in the active mode will be somewhat larger than this because (1.115) assumes $r_{ds} \cong 1/\lambda I_D$ whereas in fact $r_{ds} \cong 1/\lambda I_{D,sat}$ which is somewhat larger.

This reveals two important general conclusions about analog design: first, that to maximize dc gain, transistors should be operated with small values of V_{eff}; second, since λ is inversely proportional to gate length L, as shown in equation (1.87), intrinsic gain is maximized by taking long gate lengths.

EXAMPLE 1.15

Calculate the intrinsic gain of the transistor in Example 1.10 when in active mode.

Solution

The intrinsic gain is obtained by substituting the values for g_m and r_{ds} calculated in Example 1.10 into (1.114).

$$A_i = g_m r_{ds} = 32.9$$

This is the largest voltage gain this single transistor can achieve under these operating bias conditions.

The *unity-gain frequency* of a transistor, f_t, is intended to provide some measure of the maximum operating frequency at which a transistor might prove useful. It is the most common (though not the only) measure of transistor intrinsic speed. As with intrinsic gain, f_t is measured in the common-source configuration because of its broad relevance to both analog and digital design. An idealized measurement setup is shown in Fig. 1.27(*a*). Bias

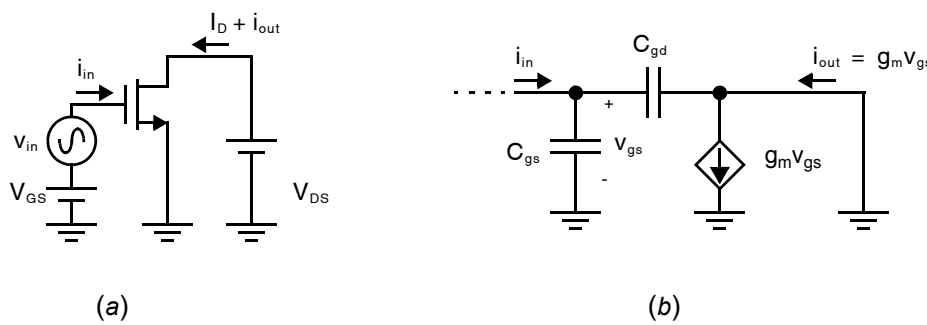


Fig. 1.27 The circuit used to characterize transistor intrinsic speed: (*a*) complete circuit; (*b*) high-frequency small-signal equivalent.

voltages V_{GS} and V_{DS} are applied to the gate and drain, respectively. A sinusoidal signal, v_{in}, with very small amplitude is superimposed on the gate voltage. This gives rise to small sinusoidal currents: i_{in} which charges and discharges the gate capacitances in response to v_{in}; and i_{out} which is a small signal current superimposed on the dc drain current, I_D. The unity-gain frequency is defined as the maximum frequency at which the amplitude of the output small-signal current, $|i_{out}|$, exceeds that of the input current, $|i_{in}|$. Although conceptually simple, this measurement is notoriously difficult to perform in the lab since the frequencies involved are high and any parasitic capacitances in the measurement setup will influence the result.

A first-order analytical solution for the transistor f_t follows straightforwardly from an analysis of the small-signal equivalent circuit in Fig. 1.27(*b*), which is left as an exercise for the reader.

$$f_t \approx \frac{g_m}{2\pi(C_{gs} + C_{gd})} \tag{1.116}$$

Although the accuracy of (1.116) is limited, we will adopt this expression as our definition of f_t because it is compact and closely related to the bandwidth of many important circuits we shall study.

Clearly, transistor intrinsic speed is dependent on biasing. Substituting (1.74) and (1.89) into (1.116) and assuming $C_{gs} \gg C_{gd}$ gives the following approximate result for the unity-gain frequency of a transistor:

$$f_t \approx \frac{\mu_n C_{ox}(W/L)V_{eff}}{2\pi C_{ox}W(2/3)L} = \frac{3\mu_n V_{eff}}{4\pi L^2} \tag{1.117}$$

Key Point: *For operation with high gain, transistors should have long gate lengths and be biased with low* V_{eff}. *For high speed, the opposite is desirable: small* L *and high* V_{eff}.

As with intrinsic gain, some important fundamental conclusions about analog design are revealed by (1.117). First, for operation at high-speed, device parasitic capacitances should be minimized, which implies minimizing the device gate length, L. Second, speed is maximized by biasing with high values of V_{eff}. These requirements are in direct conflict with those outlined for maximizing transistor intrinsic gain.

1.3 DEVICE MODEL SUMMARY

As a useful aid, all of the equations for the large-signal and small-signal modelling of diodes and MOS transistors, along with values for the various constants, are listed in the next few pages.

1.3.1 Constants

$q = 1.602 \times 10^{-19}$ C	$k = 1.38 \times 10^{-23}$ JK^{-1}
$n_i = 1.1 \times 10^{16}$ carriers/m^3 at $T = 300$ °K	$\varepsilon_0 = 8.854 \times 10^{-12}$ F/m
$K_{ox} \cong 3.9$	$K_s \cong 11.8$

1.3.2 Diode Equations

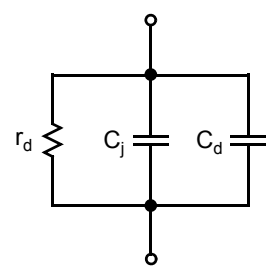
Reverse-Biased Diode (Abrupt Junction)

$C_j = \dfrac{C_{j0}}{\sqrt{1 + \dfrac{V_R}{\Phi_0}}}$	$Q = 2C_{j0}\Phi_0\sqrt{1 + \dfrac{V_R}{\Phi_0}}$
$C_{j0} = \sqrt{\dfrac{qK_s\varepsilon_0}{2\Phi_0}\dfrac{N_D N_A}{N_A + N_D}}$	$C_{j0} = \sqrt{\dfrac{qK_s\varepsilon_0 N_D}{2\Phi_0}}$ if $N_A >> N_D$
$\Phi_0 = \dfrac{kT}{q}\ln\left(\dfrac{N_A N_D}{n_i^2}\right)$	

Forward-Biased Diode

$I_D = I_S e^{V_D/V_T}$	$I_S = A_D q n_i^2\left(\dfrac{D_n}{L_n N_A} + \dfrac{D_p}{L_p N_D}\right)$
$V_T = \dfrac{kT}{q} \cong 26 \text{ mV at } 300 \text{ °K}$	

Small-Signal Model of Forward-Biased Diode



$r_d = \dfrac{V_T}{I_D}$	$C_T = C_d + C_j$
$C_d = \tau_T\dfrac{I_D}{V_T}$	$C_j \cong 2C_{j0}$
$\tau_T = \dfrac{L_n^2}{D_n}$	

1.3.3 MOS Transistor Equations

The following equations are for n-channel devices—for p-channel devices, put negative signs in front of all voltages. These equations do not account for short-channel effects (i.e., $L < 2L_{min}$).

Triode Region ($V_{GS} > V_{tn}$, $V_{DS} \le V_{eff}$)

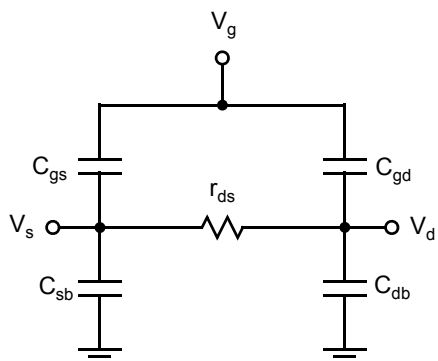
$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right]$$

$$V_{eff} = V_{GS} - V_{tn} \qquad V_{tn} = V_{tn\text{-}0} + \gamma(\sqrt{V_{SB} + 2\phi_F} - \sqrt{2\phi_F})$$

$$\phi_F = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) \qquad \gamma = \frac{\sqrt{2qK_{si}\varepsilon_0 N_A}}{C_{ox}}$$

$$C_{ox} = \frac{K_{ox}\varepsilon_0}{t_{ox}}$$

Small-Signal Model in Triode Region (for $V_{DS} << V_{eff}$)



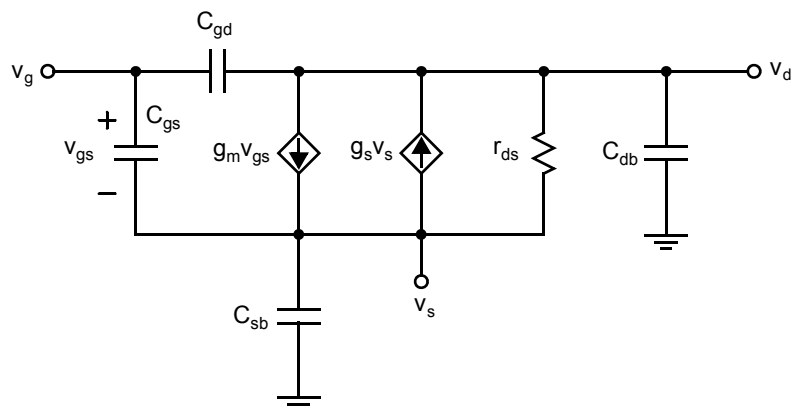
$$r_{ds} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}}$$

$$C_{gd} = C_{gs} \cong \frac{1}{2}WLC_{ox} + WL_{ov}C_{ox} \qquad C_{sb} = C_{db} = \frac{C_{j0}(A_s + WL/2)}{\sqrt{1 + \frac{V_{sb}}{\Phi_0}}}$$

Active (or Pinch-Off) Region ($V_{GS} > V_{tn}$, $V_{DS} \geq V_{eff}$)

$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS}-V_{tn})^2[1+\lambda(V_{DS}-V_{eff})]$	
$\lambda \propto \frac{1}{L\sqrt{V_{DS}-V_{eff}+\Phi_0}}$	$V_{tn} = V_{tn\text{-}0} + \gamma(\sqrt{V_{SB}+2\phi_F}-\sqrt{2\phi_F})$
$V_{eff} = V_{GS}-V_{tn} = \sqrt{\frac{2I_D}{\mu_n C_{ox}W/L}}$	

Small-Signal Model (Active Region)



$g_m = \mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}$	$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$
$g_m = \frac{2I_D}{V_{eff}}$	$g_s = \frac{\gamma g_m}{2\sqrt{V_{SB}+\lvert 2\phi_F\rvert}}$
$r_{ds} = \frac{1}{\lambda I_{D\text{-}sat}} \cong \frac{1}{\lambda I_D}$	$g_s \cong 0.2 g_m$
$\lambda = \frac{k_{rds}}{2L\sqrt{V_{DS}-V_{eff}+\Phi_0}}$	$k_{rds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}}$
$C_{gs} = \frac{2}{3}WLC_{ox} + WL_{ov}C_{ox}$	$C_{gd} = WL_{ov}C_{ox}$

$C_{sb} = (A_s + WL)C_{js} + P_sC_{j\text{-}sw}$	$C_{js} = \dfrac{C_{j0}}{\sqrt{1 + V_{SB}/\Phi_0}}$
$C_{db} = A_dC_{jd} + P_dC_{j\text{-}sw}$	$C_{jd} = \dfrac{C_{j0}}{\sqrt{1 + V_{DB}/\Phi_0}}$

1.4 ADVANCED MOS MODELLING

The square-law relationship between effective gate-source voltage and drain current expressed in equation (1.67) is only valid for active operation in strong inversion. For very small (and negative) values of V_{eff}, MOS devices operate in weak inversion, also called subthreshold operation, where an exponential voltage-current relationship holds. For large values of V_{eff}, a combination of effects give rise to a sub-square-law voltage–current relationship. Both are considered in this section. Other advanced modelling concepts that an analog microcircuit designer is likely to encounter are also covered, including parasitic resistances, short channel effects, and leakage currents.

1.4.1 Subthreshold Operation

Key Point: *In subthreshold operation, also called weak inversion, transistors obey an exponential voltage–current relationship instead of a square-law. Hence, a small but finite current flows even when* $V_{GS} = 0$.

The device equations presented for active or triode region MOS transistors in the preceding sections are all based on the assumption that V_{eff} is greater than about 100 mV and the device is in strong inversion. When this is not the case, the accuracy of the square-law equations is poor. For negative values of V_{eff}, the transistor is in weak inversion and is said to be operating in the *subthreshold region.* In this regime, the dominant physical mechanism for conduction between drain and source is diffusion, not drift as in strong inversion, and the transistor is more accurately modelled by an exponential relationship between its control voltage (at the gate) and current, somewhat similar to a bipolar transistor. In the subthreshold region, the drain current is approximately given by an exponential relationship.

$$I_{D(sub-th)} \cong I_{D0}\left(\frac{W}{L}\right)e^{(qV_{eff}/nkT)} \tag{1.118}$$

where

$$n = \frac{C_{ox} + C_{j0}}{C_{ox}} \approx 1.5 \tag{1.119}$$

$$I_{D0} = (n-1)\mu_nC_{ox}\left(\frac{kT}{q}\right)^2 \tag{1.120}$$

and it has been assumed that $V_S = 0$ and $V_{DS} > 75$ mV. Not captured here is drain-induced barrier lowering, a short-channel effect that causes subthreshold current to also depend on drain-source voltage.

Plotting drain current on a logarithmic axis versus V_{GS} in the subthreshold regime gives a straight line. The inverse of this slope, called the *subthreshold slope* and equal to $\ln(10)\cdot nkT/q = 2.3nkT/q$, is a measure of the voltage change in V_{GS} required to effect an order-of-magnitude change in subthreshold drain current.

Also note that the current does not drop to zero even when $V_{GS} = 0V$. This residual drain current is called *subthreshold leakage* and considering equation (1.118) is given by

$$I_{off} = I_{D0}\left(\frac{W}{L}\right)e^{(-qV_t/nkT)} = (n-1)\mu_n C_{ox}\left(\frac{W}{L}\right)\left(\frac{kT}{q}\right)^2 e^{(-qV_t/nkT)} \tag{1.121}$$

Equation (1.121) reveals a complex temperature dependence due to the explicit inclusion of absolute temperature (T), and due to the temperature dependence of carrier mobility (μ_n) and threshold voltage (V_t). In general, subthreshold leakage increases significantly with temperature and is often a dominant source of power consumption in analog circuits that are temporarily powered down or in standby.

EXAMPLE 1.16

A transistor has a subthreshold leakage current of $I_{off} = 10$ nA at 300 K. How much must V_{eff} decrease to reduce I_{off} to 1 nA?

Solution

From (1.118), it is evident that to decrease subthreshold drain current by a factor of 10 requires that V_{eff} decrease by $\ln(10)\cdot nkT/q \cong 90$ mV at room temperature. This can be achieved either by decreasing V_{GS}, or increasing V_{tn}.

The transconductance of a transistor in the subthreshold region is determined by taking the derivative of (1.118) with respect to V_{GS} resulting in

$$g_{m(sub-th)} = \left(\frac{q}{nkT}\right)I_{D0}\left(\frac{W}{L}\right)e^{(qV_{eff}/nkT)} = \frac{qI_D}{nkT} \tag{1.122}$$

Note that for fixed drain current, I_D, the transconductance in subthreshold is independent of V_{eff}. Normalizing the transconductance with respect to the drain current yields a constant value in subthreshold,

$$\frac{g_{m(sub-th)}}{I_{D(sub-th)}} = \frac{q}{nkT} \tag{1.123}$$

As V_{GS} is increased well beyond V_t, the MOS device enters strong inversion and the transconductance then decreases in inverse proportion to V_{eff}, as indicated by (1.78). Hence, for a fixed drain current, transconductance of a MOS device is maximized in the subthreshold region with the value given in (1.122). In other words, a targeted value of transconductance can be achieved with less drain current in the subthreshold region than in strong inversion. However, to achieve practically useful values of transconductance in the subthreshold region requires a transistor with a very large aspect ratio, W/L. Such large aspect ratios generally imply large parasitic capacitances making very high-speed operation difficult. Therefore, subthreshold operation is useful primarily when speed can be sacrificed for lower drain currents and, hence, generally lower power consumption.

Key Point: *For a fixed drain current, the small-signal transconductance of a MOS device is maximized in subthreshold operation at a value of* $g_m = qI_D/nkT$.

The transition between subthreshold and strong inversion is not abrupt. For a broad range of gate-source voltages, both diffusion and drift currents with comparable magnitudes exist making accurate device modeling in this region notoriously difficult. This is generally referred to as *moderate inversion*. To get a feel for what gate-source voltages imply moderate inversion, equate the expressions for g_m in active mode (1.78) and $g_{m(sub-th)}$ (1.122) and solve for V_{eff} resulting in

$$V_{eff} = \frac{2nkT}{q} \tag{1.124}$$

At room temperature and assuming $n \approx 1.5$, moderate inversion occurs around $V_{eff} \approx 75\text{mV}$, increasing to approximately 90 mV at 350 K. Hence, to ensure operation in strong inversion, prudent designers generally take $V_{eff} > 100$ mV.

EXAMPLE 1.17

Estimate the value of n from the data for the NMOS device in Example 1.9 at $T = 300\text{K}$.

Solution

The logarithmic plot of I_D versus V_{GS} in Fig. 1.28(*a*) has a slope equal to $(2.3nkT)/q$ which in this case can be used to estimate n. The dashed line shows a reasonable fit with $n = 1.6$. This estimate can be confirmed by looking at the plot of g_m/I_D also on Fig. 1.28(*b*), which should reach a maximum of q/nkT in subthreshold. Again, the dashed line with $n = 1.6$ is a reasonable approximation. Also shown on that plot is the result for the square-law model based on (1.78) based on the threshold voltage of $V_{tn} \cong 0.45$ V obtained in Example 1.9. Note that the two regions intersect at $V_{GS} = V_{tn} + 2nkT/q$.

1.4.2 Mobility Degradation

Transistors subjected to large electric fields experience a degradation in the effective mobility of their carriers. Large lateral electric fields, as shown in Fig. 1.29 for an NMOS device, accelerate carriers in the channel up to a

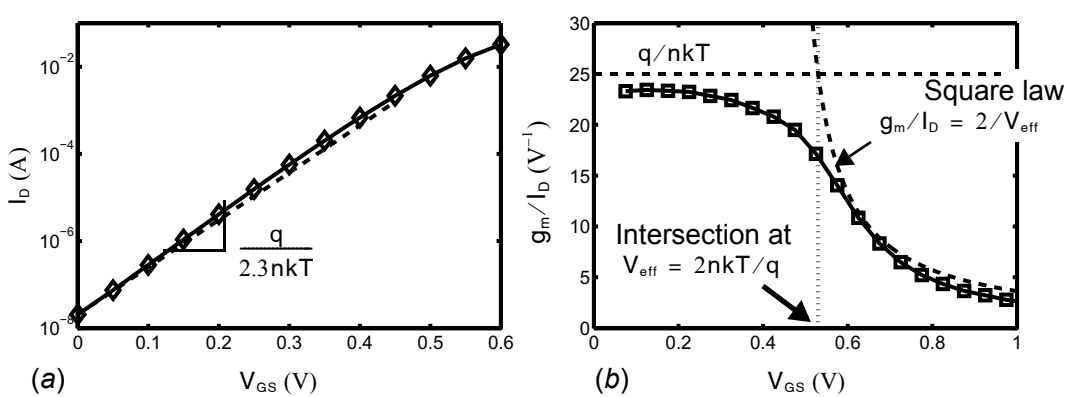


Fig. 1.28 Drain current and transconductance of a MOS device in subthreshold.

maximum velocity, approximately 10^7 cm/s in silicon. This is referred to as *velocity saturation*. Vertical electric fields greater than approximately $5 \cdot 10^6$ V/m cause the effective channel depth to decrease and also cause more charge-carrier collisions.

For the purposes of design, these effects may be modeled together by an effective carrier mobility that decreases at high V_{eff},

$$\mu_{n,\,eff} \cong \frac{\mu_n}{([1 + (\theta V_{eff})^m])^{1/m}} \tag{1.125}$$

where θ and m are device parameters. Substituting the effective carrier mobility, $\mu_{n,\,eff}$, from (1.125) in place of μ_n in equation (1.72) gives a new expression for the drain current incorporating mobility degradation.

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{eff}^2\left(\frac{1}{[1 + (\theta V_{eff})^m]^{1/m}}\right) \tag{1.126}$$

For small values of V_{eff}, the final term in equation (1.126) approaches unity and the entire expression simplifies to the square-law relationship in (1.72). For values of $V_{eff} \gg 1/\theta$, the final term in (1.126) approaches $(1/\theta V_{eff})$ resulting in a linear relationship between effective gate-source voltage and current. Taking the derivative of (1.126) with respect to gate-source voltage assuming large V_{eff} gives a small-signal transconductance that is independent of drain current,

$$g_{m(\text{mob-deg})} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\frac{1}{\theta} \tag{1.127}$$

Equation (1.127) establishes the maximum transconductance achievable with a given transistor. A square-law model for transconductance (1.73) predicts a transconductance equal to (1.127) when $V_{eff} = 1/2\theta$. Due to mobility degradation, increases in V_{eff} beyond $1/2\theta$:

- fail to provide significant increases in small-signal transconductance,
- reduce the available signal swing limited by the fixed supply voltages, and
- reduce transistor intrinsic gain A_i dramatically

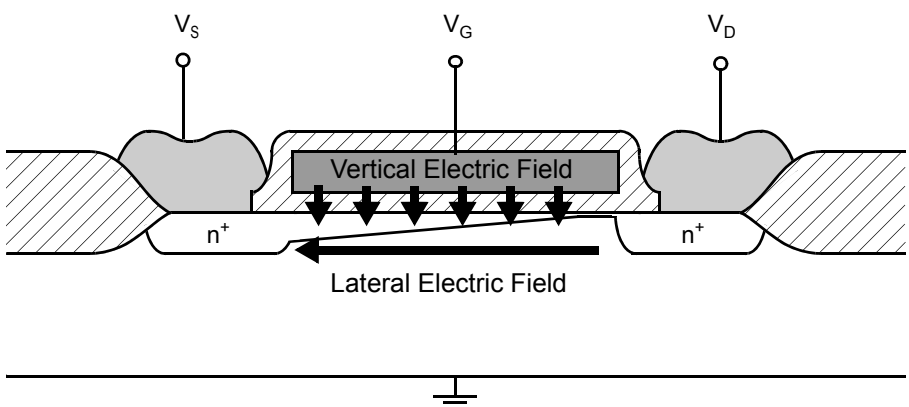


Fig. 1.29 An NMOS device in active mode (saturation) identifying the lateral and vertical electric field components.

For these reasons, operation under mobility degradation is generally avoided in analog design, although this is becoming increasingly difficult in modern CMOS processes. Transistors are, however, sometimes biased with $V_{eff} > 1/2\theta$ when their size must be limited. For example, when very-high frequency operation is sought, parasitic capacitances must be minimized by keeping all transistor dimensions as small as possible.

Key Point: *For large values of V_{eff}, transistors have a sub-square-law voltage current relationship, tending towards a linear relationship for very large values of V_{eff}.*

For operating conditions around $V_{eff} \approx 1/2\theta$, more sophisticated modeling is required to obtain high accuracy [Sodini, 1984]. However, a piecewise-linear assumption whereby (1.73) is applied when $V_{eff} < 1/2\theta$ and (1.127) when $V_{eff} > 1/2\theta$ is often sufficient for first pass design.

In the past, mobility degradation appeared only at extremely large values of V_{eff}, but this is no longer always the case. Shrinking MOS device dimensions have meant that lower voltages are required to generate the critical electric fields. The value of θ increases from around 0.06 V^{-1} for a 0.8-μm-long transistor in a 0.8-μm CMOS process, to greater than 2 V^{-1} for minimum-sized devices in modern CMOS processes. Hence, mobility degradation can appear at V_{eff} values of only a few hundred mV. Since (1.124) places a fundamental lower limit of $V_{eff} > 100$ mV to ensure operation in strong inversion there remains only a narrow range of effective gate-source voltages for which the square-law voltage–current relationship expressed in (1.72) remains valid in nanoscale CMOS devices. In many voltage-to-current conversion circuits that rely on the square-law characteristic, this inaccuracy can be a major source of error. Taking channel lengths larger than the minimum allowed helps to minimize this degradation.

EXAMPLE 1.18

Estimate the value of θ based on the data in Fig. 1.30.

Solution

Mobility degradation is evident at high values of V_{eff} where the drain current in Fig. 1.30(*a*) is clearly subquadratic. An estimate of θ is more easily obtained from the plot of g_m versus V_{GS} in Fig. 1.30(*b*). The value of V_{eff} at which the transconductance ceases to follow a linear relationship is approximately $1/2\theta$. In this case, this

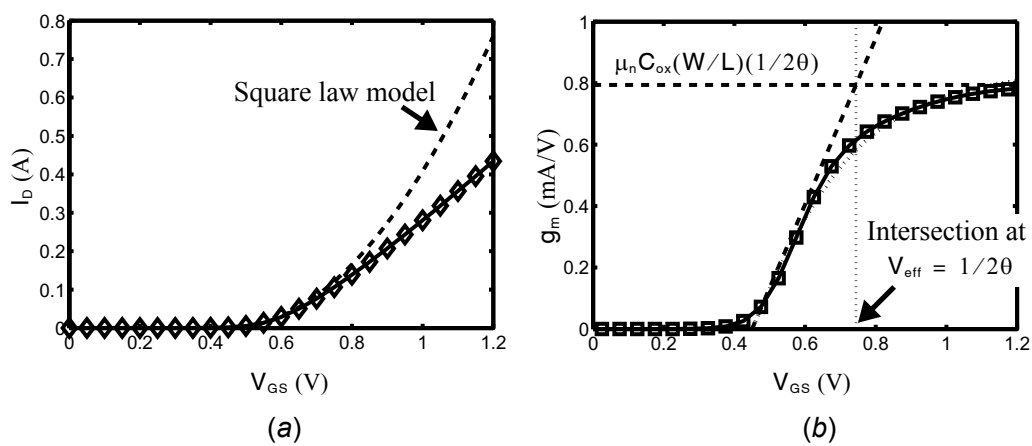


Fig. 1.30 Drain current and transconductance of a MOS device in at large values of V_{eff}.

occurs at approximately $V_{GS} = 0.75$ V or $V_{eff} = 0.3$ V, corresponding to $\theta \cong 1.7\ V^{-1}$. At $V_{eff} > 0.3$ V, the constant $\mu_n C_{ox}(W/L)(1/2\theta)$ can be used as a rough estimate of g_m. More accuracy is provided using the model of (1.126) with, in this case, a value $m = 1.6$ resulting in the dotted line on Fig. 1.30(*b*).

1.4.3 Summary of Subthreshold and Mobility Degradation Equations

The key expressions for operation in subthreshold, strong-inversion, and mobility degradation are summarized in Table 1.1 and the progression of transistor drain current and small-signal transconductance as its gate-source voltage is swept is sketched in Fig. 1.31. Transition regions appear at the borders between these operating modes where the voltage-current relationship and small-signal transconductance are compromises between the expressions in the table. Generally, operation in or near subthreshold is considered only for analog circuits where low-power but low-speed operation is required. Operation under mobility degradation is necessary only when very high-speed is required, thus demanding minimal parasitic capacitances and, hence, low device aspect ratios.

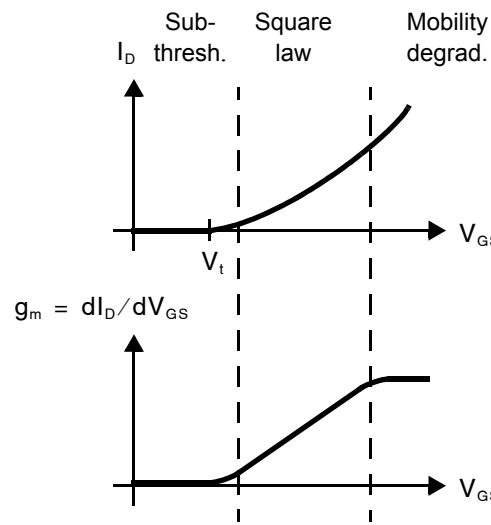


Fig. 1.31 The progression of transistor drain current and small-signal transconductance from subthreshold to square law to mobility degradation.

1.4.4 Parasitic Resistances

Parasitic resistance appears in series with all four MOSFET terminals. The resistance of the polysilicon gate carries no dc current, but may be significant in small-signal analysis since it is in series with the gate-source capacitance, C_{gs}. Contacts to the drain and source region are resistive and may have significant voltage drops across them when carrying large currents. Additionally, in modern CMOS processes, the drain

Table 1.1 **A summary of MOS device operation in the subthreshold, strong inversion, and mobility degradation regions.**

	Subthreshold (exponential)	Strong Inversion (square-law)	Mobility Degradation (linear)
Region of validity	$V_{eff} \lesssim 0$	$\frac{2nkT}{q} < V_{eff} < \frac{1}{2\theta}$	$V_{eff} > \frac{1}{2\theta}$
Drain current, I_D	$I_{D0}\left(\frac{W}{L}\right)e^{(qV_{eff}/nkT)}$	$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{eff}^2$	$\frac{0.5\mu_n C_{ox}(W/L)V_{eff}^2}{[1+(\theta V_{eff})^m]^{1/m}}$
Small-signal transconductance, g_m	$\frac{qI_D}{nkT}$	$\frac{2I_D}{V_{eff}} = \mu_n C_{ox}\frac{W}{L}V_{eff}$	$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\frac{1}{\theta}$
Most useful	Very low-power operation	Most analog design	Very high-speed operation

and source regions themselves are engineered to have a very shallow depth near the channel region. These shallow source and drain extensions present a narrow cross-sectional area through which all channel current must flow resulting in significant series resistance. Finally, the body terminal is relatively lightly doped semiconductor and, hence, may impose a significant series resistance. All of these parasitics are generally on the order of a few Ohms, and hence are only considered when they are conducting very large currents, or at very high frequencies where the resistances combine with transistor parasitic capacitances to form small-signal poles of significance. Furthermore, they can often be minimized by proper design (for example, by providing additional contacts). They are considered negligible throughout the remainder of the text.

1.4.5 Short-Channel Effects

A number of short-channel effects degrade the operation of MOS transistors as device dimensions are scaled down. These effects include reduced output impedance and hot-carrier effects (such as oxide trapping and substrate currents). These short-channel effects will be briefly described here. For more detailed modelling of short-channel effects, see [Wolf, 1995].

Transistors with short channel lengths experience a reduced output impedance because depletion region variations at the drain end (which affect the effective channel length) have an increased proportional effect on the drain current. In addition, a phenomenon known as drain-induced barrier lowering (DIBL) effectively lowers V_t as V_{DS} is increased, thereby further lowering the output impedance of a short-channel device.

Another important short-channel effect is due to *hot carriers*. These high-velocity carriers can cause harmful effects, such as the generation of electron-hole pairs by impact ionization and avalanching. These extra electron-hole pairs can cause currents to flow from the drain to the substrate, as shown in Fig. 1.32. This effect can be modelled by a finite drain-to-ground impedance. As a result, this effect is one of the major limitations on achieving very high output impedances of cascode current sources. In addition, this current flow can cause voltage drops across the substrate and possibly cause latch-up, as the next chapter describes.

Another hot-carrier effect occurs when electrons gain energies high enough to tunnel into and possibly through the thin gate oxide. Thus, this effect can cause dc gate currents. However, often more harmful is the fact that any charge trapped in the oxide will cause a shift in transistor threshold voltage. As a result, hot carriers are one of the major factors limiting the long-term reliability of MOS transistors.

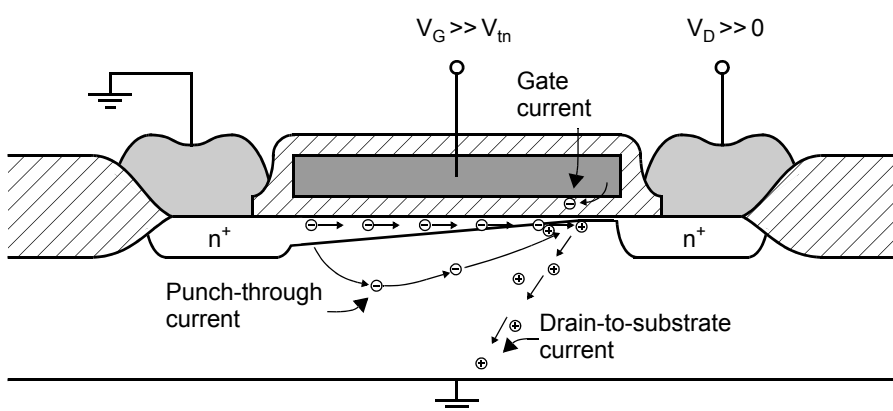


Fig. 1.32 An illustration of hot carrier effects in an n-channel MOS transistor. Drain-to-substrate current is caused by electron-hole pairs generated by impact ionization at the drain end of the channel.

A third hot-carrier effect occurs when electrons with enough energy *punch through* from the source to the drain. As a result, these high-energy electrons are no longer limited by the drift equations governing normal conduction along the channel. This mechanism is somewhat similar to punch-through in a bipolar transistor, where the collector depletion region extends right through the base region to the emitter. In a MOS transistor, the channel length becomes effectively zero, resulting in unlimited current flow (except for the series source and drain impedances, as well as external circuitry). This effect is an additional cause of lower output impedance and possibly transistor breakdown.

All of the hot-carrier effects just described are more pronounced for n-channel transistors than for their p-channel counterparts because electrons have larger velocities than holes.

1.4.6 Leakage Currents

Leakage currents impose limitations on how long a dynamically charged signal can be maintained in a high impedance state, and on the minimum power consumption achievable when an analog circuit is in an idle (power-down) state. Three leakage currents are important in MOS transistors: subthreshold leakage, gate leakage, and junction leakage. Of these, subthreshold leakage is often the largest. It results in a finite drain current even when the transistor is off, and was described in detail in section 1.4.1. Gate leakage results from quantum-mechanical tunneling of electrons through very thin gate oxides, and can be significant in digital circuits and when large gate areas are used, for example to implement a capacitor. Finally, the reverse-biased source-body and drain-body pn junctions conduct a finite leakage current. This leakage can be important, for example, in estimating the maximum time a sample-and-hold circuit or a dynamic memory cell can be left in hold mode. The leakage current of a reverse-biased junction (not close to breakdown) is approximately given by

$$I_{lk} \cong \frac{qA_j n_i}{2\tau_0} x_d \tag{1.128}$$

where A_j is the junction area, n_i is the intrinsic concentration of carriers in undoped silicon, τ_0 is the effective minority carrier lifetime, x_d is the thickness of the depletion region, and τ_0 is given by

$$\tau_0 \cong \frac{1}{2}(\tau_n + \tau_p) \tag{1.129}$$

where τ_n and τ_p are the electron and hole lifetimes. Also, x_d is given by

$$x_d = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}(\Phi_0 + V_r)} \tag{1.130}$$

and n_i is given by

$$n_i \cong \sqrt{N_C N_V}\ e^{(-E_g)/(kT)} \tag{1.131}$$

where N_C and N_V are the densities of states in the conduction and valence bands and E_g is the difference in energy between the two bands.

Since the intrinsic concentration, n_i, is a strong function of temperature (it approximately doubles for every temperature increase of 11 °C for silicon), the leakage current is also a strong function of temperature. The leakage current roughly doubles for every 11 °C rise in temperature; thus, at higher temperatures it is much larger than at room temperature.

1.5 SPICE MODELLING PARAMETERS

This section briefly describes some of the important model parameters for diodes and MOS transistors used during a SPICE simulation. It should be noted here that not all SPICE model parameters are described. However, enough are described to enable the reader to understand the relationship between the relative parameters and the corresponding constants used when doing hand analysis.

1.5.1 Diode Model

There are a number of important dc parameters. The constant I_S is specified using either the parameter IS or JS in SPICE. These two parameters are synonyms, and only one should be specified. A typical value specified for I_S might be between 10^{-18} A and 10^{-15} A for small diodes in a microcircuit. Another important parameter is called the *emission coefficient,* n_j. This constant multiplies V_T in the exponential diode I-V relationship given by

$$I_D = I_S e^{V_{BE}/(n_j V_T)} \tag{1.132}$$

The SPICE parameter for n_j is N and is defaulted to 1 when not specified (1 is a reasonable value for junctions in an integrated circuit). A third important dc characteristic is the *series resistance,* which is specified in SPICE using RS. It should be noted here that some SPICE programs allow the user to specify the area of the diode, whereas others expect absolute parameters that already take into account the effective area. The manual for the program being used should be consulted.

The diode transit time is specified using the SPICE parameter TT. The most important capacitance parameter specified is CJ. CJO and CJ are synonyms—one should never specify both. This parameter specifies the capacitance at 0-V bias. Once again, it may be specified as absolute or as relative to the area (i.e., F/m^2), depending on the version of SPICE used. Also, the area junction grading coefficient, MJ, might be specified to determine the exponent used in the capacitance equation. Typical values are 0.5 for abrupt junctions and 0.33 for graded junctions. In some SPICE versions, it might also be possible to specify the sidewall capacitance at 0-V bias as well as its grading junction coefficient. Finally, the built-in potential of the junction, which is also used in calculating the capacitance, can be specified using PB. PHI, VJ, and PHA are all synonyms of PB.

Reasonably accurate diode simulations can usually be obtained by specifying only IS, CJ, MJ, and PB. However, most modern versions of SPICE have many more parameters that can be specified if one wants accurate temperature and noise simulations. Users should consult their manuals for more information.

Table 1.2 summarizes some of the more important diode parameters. This set of parameters constitutes a minimal set for reasonable simulation accuracy under ordinary conditions.

Table 1.2 Important SPICE parameters for modelling diodes.

SPICE Parameter	Model Constant	Brief Description	Typical Value
IS	I_S	Transport saturation current	10^{-17} A
RS	R_d	Series resistance	30 Ω
TT	τ_T	Diode transit time	12 ps
CJ	C_{j0}	Capacitance at 0-V bias	0.01 pF
MJ	m_j	Diode grading coefficient exponent	0.5
PB	Φ_0	Built-in diode contact potential	0.9 V

1.5.2 MOS Transistors

Modern MOS models are quite complicated, so only some of the more important MOS parameters used in SPICE simulations are described here. These parameters are used in what are called the Level 2 or Level 3 models. The model level can be chosen by setting the SPICE parameter LEVEL to either 2 or 3. The oxide thickness, t_{ox}, is specified using the SPICE parameter TOX. If it is specified, then it is not necessary to specify the thin gate-oxide capacitance (C_{ox}, specified by parameter COX). The mobility, μ_n, can be specified using UO. If UO is specified, the intrinsic transistor conductance ($\mu_n C_{ox}$) will be calculated automatically, unless this automatic calculation is overridden by specifying either KP (or its synonym, BETA). The transistor threshold voltage at $V_S = 0$ V, V_{tn}, is specified by VTO. The body-effect parameter, γ, can be specified using GAMMA, or it will be automatically calculated if the substrate doping, N_A, is specified using NSUB. Normally, one would not want SPICE to calculate γ because the effective substrate doping under the channel can differ significantly from the substrate doping in the bulk due to threshold-voltage adjust implants. The output impedance constant, λ, can be specified using LAMBDA. Normally, LAMBDA should not be specified since it takes precedence over internal calculations and does not change the output impedance as a function of different transistor lengths or bias voltages (which should be the case). Indeed, modelling the transistor output impedance is one of weakest points in SPICE. If LAMBDA is not specified, it is calculated automatically. The surface inversion potential, $|2\phi_F|$, can be specified using PHI, or it will be calculated automatically. Another parameter usually specified is the lateral diffusion of the junctions under the gate, L_D, which is specified by LD. For accurate simulations, one might also specify the resistances in series with the source and drain by specifying RS and RD (typically only the source resistance is important). Many other parameters exist to model such things as short-channel effects, subthreshold effects, and channel-width effects, but these parameters are outside the scope of this book.

The modelling of parasitic capacitances in SPICE is quite involved. The capacitances under the junctions per unit area at 0-V bias, (i.e., C_{j0}) can be specified using CJ or can be calculated automatically from the specified substrate doping. The sidewall capacitances at 0 V, $C_{j\text{-}sw0}$, should normally be specified using CJSW because this parameter is used to calculate significant parasitic capacitances. The bulk grading coefficient specified by MJ can usually be defaulted to 0.5. Similarly, the sidewall grading coefficient specified by MJSW can usually be defaulted to 0.33 (SPICE assumes a graded junction). The built-in bulk-to-junction contact potential, Φ_0, can be specified using PB or defaulted to 0.8 V (note that 0.9 V would typically be more accurate, but the resulting simulation differences are small). Sometimes the gate-to-source or drain-overlap capacitances can be specified using CGSO or CGDO, but normally these would be left to be calculated automatically using COX and LD.

Some of the more important parameters that should result in reasonable simulations (except for modelling short-channel effects) are summarized in Table 1.3 for both n- and p-channel transistors. Table 1.3 lists reasonable parameters for a typical 0.8-μm technology.

1.5.3 Advanced SPICE Models of MOS Transistors

Although the SPICE model parameters presented in the last section provide reasonable accuracy for long-channel devices, for channel lengths $L \ll 1\mu m$ their accuracy becomes very poor. Many SPICE MOS models have therefore been developed to try to capture higher-order effects. A summary of the capabilities of the more common modern SPICE model formats is provided in Table 1.4.

Unfortunately, these SPICE models require over 100 parameters to accurately capture transistor operation in all of its modes over a wide range of temperatures. Many parameters are required because, for very small device sizes, fundamental constants such as threshold voltage and effective carrier mobility become dependent on the transistor's exact width and length. Hence, it is strongly recommended that analog designers use only a small set of "unit-sized" transistors, forming all transistors from parallel combinations of these elementary devices.

Table 1.3 A reasonable set of Level 2 or 3 MOS parameters for a typical 0.8-μm technology.

SPICE Parameter	Model Constant	Brief Description	Typical Value
VTO	V_{tn}:V_{tp}	Transistor threshold voltage (in V)	0.8:–0.9
UO	μ_n:μ_p	Carrier mobility in bulk (in $cm^2/V{\cdot}s$)	500:175
TOX	t_{ox}	Thickness of gate oxide (in m)	1.8×10^{-8}
LD	L_D	Lateral diffusion of junction under gate (in m)	6×10^{-8}
GAMMA	γ	Body-effect parameter	0.5: 0.8
NSUB	N_A:N_D	The substrate doping (in cm^{-3})	3×10^{16}:7.5×10^{16}
PHI	$\lvert 2\phi_F \rvert$	Surface inversion potential (in V)	0.7
PB	Φ_0	Built-in contact potential of junction to bulk (in V)	0.9
CJ	C_{j0}	Junction-depletion capacitance at 0-V bias (in F/m^2)	2.5×10^{-4}:4.0×10^{-4}
CJSW	$C_{j\text{-}sw0}$	Sidewall capacitance at 0-V bias (in F/m)	2.0×10^{-10}:2.8×10^{-10}
MJ	m_j	Bulk-to-junction exponent (grading coefficient)	0.5
MJSW	$m_{j\text{-}sw}$	Sidewall-to-junction exponent (grading coefficient)	0.3

Table 1.4 A summary of modern SPICE model formats.

SPICE Model	Main strengths compared with previous device models
BSIM3	Improved modeling of moderate inversion, and the geometry-dependence of device parameters. This also marked a return to a more physics-based model as opposed to the preceding highly empirical models.
EKV	Relates terminal currents and voltages with unified equations that cover all modes of transistor operation, hence avoiding discontinuities at transitions between, for example, weak and strong inversion. Also handles geometry-dependent device parameters.
BSIM4	Improved modeling of leakage currents and short-channel effects, noise, and parasitic resistance in the MOSFET terminals, as well as continued improvements in capturing the geometry-dependence of device parameters.
PSP	Improved modeling of noise and the many short-channel and layout-dependent effects now dominant in nanoscale CMOS devices. Particular effort was made to accurately model nonlinearities, which requires accuracy in the high-order derivatives of the transistor's voltage–current relationships.

For example, for a minimum gate-length of L_{min}, device sizes with $(W/L) = (8L_{min}/L_{min})$, $(12L_{min}/1.5L_{min})$, $(16L_{min}/2L_{min})$, and $(32L_{min}/4L_{min})$ in both NMOS and PMOS varieties might be chosen. Then, if a transistor of size $(W/L) = (240L_{min}/2L_{min})$ is desired, simply combine 15 of the $(16L_{min}/2L_{min})$-devices in parallel as shown in Fig. 1.33. Of course this practice restricts the device sizes available for design, but the benefits of having device parameters that are consistent and well-understood far outweigh this minor drawback.

For each device in the set, rough estimates of a few basic model parameters such as $\mu_n C_{ox}$, V_{tn}, etc. may be obtained from simulation data, or better yet measurements, of the unit-sized devices following the procedures in Examples 1.9, 1.11, 1.17, and 1.18. These rough model parameters may be used for the many quick hand-calculations performed in the course of first-pass design. Similarly, the unit-transistor's parasitic capacitances may

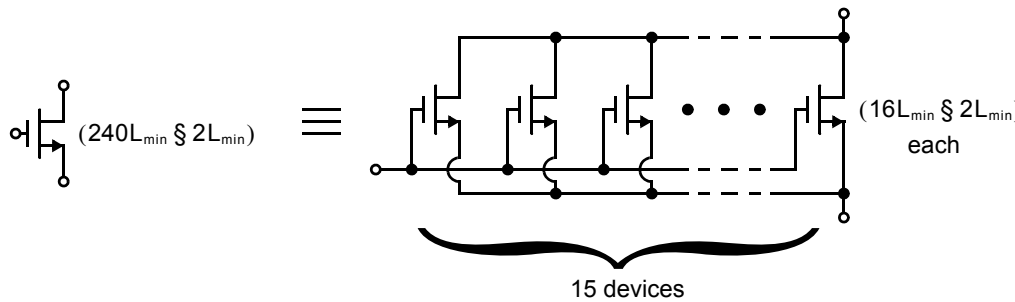


Fig. 1.33 Realization of a transistor by parallel combination of small unit-sized transistor elements.

be observed under a typical biasing condition; these capacitance values can then be used to obtain rough estimates of circuit parasitics. Several sets of such parameters that are representative of various CMOS technologies are presented in Table 1.5. When refinement and detailed verification of a design are required, these are of course performed with the aid of SPICE and the complete device models.

Some designers prefer to forgo the extraction of device parameters from data and simply refer to plots of device data throughout the design process, looking up small-signal parameters at different operating points. In any case, performing basic device simulations at the outset of any analog design is useful, not only for extracting approximate device parameters from complex MOS models, but also because it may expose shortcomings in the device models themselves or even the simulation environment. For example, discontinuities observed in plots of drain current or its derivatives are an indication that the model may be unreliable in that region.

SPICE! *MOSFET model files corresponding to each of the columns in Table 1.5 are available on the book web site for simulation exercises.*

Table 1.5 MOSFET parameters representative of various CMOS technologies and used for rough hand calculations in this text.

	0.8 μm		0.35 μm		0.18 μm		45 nm	
Technology	**NMOS**	**PMOS**	**NMOS**	**PMOS**	**NMOS**	**PMOS**	**NMOS**	**PMOS**
μC_{ox} ($\mu A/V^2$)	92	30	190	55	270	70	280	70
V_{t0} (V)	0.80	–0.90	0.57	–0.71	0.45	–0.45	0.45	–0.45
$\lambda \cdot L$ (μm/V)	0.12	0.08	0.16	0.16	0.08	0.08	0.10	0.15
C_{ox} ($fF/\mu m^2$)	1.8	1.8	4.5	4.5	8.5	8.5	25	25
t_{ox} (nm)	18	18	8	8	5	5	1.2	1.2
n	1.5	1.5	1.8	1.7	1.6	1.7	1.85	1.85
θ (1/V)	0.06	0.135	1.5	1.0	1.7	1.0	2.3	2.0
m	1.0	1.0	1.8	1.8	1.6	2.4	3.0	3.0
$C_{ov}/W = L_{ov}C_{ox}$ (fF/μm)	0.20	0.20	0.20	0.20	0.35	0.35	0.50	0.50
$C_{db}/W \approx C_{sb}/W$ (fF/μm)	0.50	0.80	0.75	1.10	0.50	0.55	0.45	0.50

1.6 PASSIVE DEVICES

Passive components are often required for analog design. However, since transistor performance is the primary consideration in the development of most integrated-circuit fabrication technologies, integrated-circuit passive components exhibit significant nonidealities that must be understood by the analog designer. The most common passive components in analog integrated-circuit design are resistors and capacitors.

1.6.1 Resistors

Strip Resistors

The simplest realization of an integrated-circuit resistor is nothing more than a strip of conductive material above the silicon substrate, as shown in Fig. 1.34. The conductivity of the material from which the strip is made is generally characterized by its sheet resistance, $R_\square$, which is derived in terms of basic material properties in Section 1.7.4 and has units of Ω. This, along with the dimensions of the strip, determine the value of the resistor.

$$R = R_\square \left(\frac{L}{W}\right) \tag{1.133}$$

This equation is also useful for calculating the resistance of interconnects used in integrated circuits. Multiple strips may be combined in series or in parallel to realize the desired total resistance without requiring impractically large or small values of L or W.

Strip resistors inevitably have parasitic capacitances to the silicon substrate which is generally held at a constant reference potential. A simple small-signal model is shown in Fig. 1.34. Although strictly speaking the capacitance is distributed evenly along the resistor, this first-order model simply splits the total capacitance into two lumped capacitors at either end of the resistor.

EXAMPLE 1.19

In many CMOS manufacturing processes, polysilicon strips are used to provide a controllable sheet resistance for analog design. A typical value is $R_\square = 500\Omega$. If each strip is $1\mu m$ wide and $5\mu m$ long, how many must be connected in series to make a resistor of value $50k\Omega$?

Solution

Each strip has a resistance of

$$R = R_\square \left(\frac{L}{W}\right) = 500\Omega\left(\frac{5\mu m}{1\mu m}\right) = 2.5k\Omega$$

Hence, 20 must be connected in series to realize the desired total resistance,

$$R_{tot} = 20R = 20(2.5k\Omega) = 50k\Omega$$

Semiconductor Resistors

Integrated-circuit manufacturing processes that are developed primarily for digital design may not include any materials with sufficient sheet resistance to realize practically useful resistor values. In these processes, a lightly-doped section of the silicon substrate with contacts at either end may be used as a resistor. The common case of a

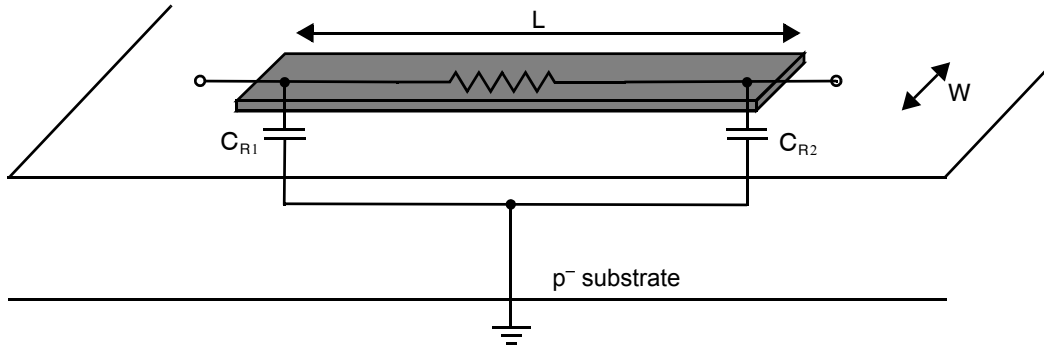


Fig. 1.34 A strip-resistor and its model including parasitic capacitances to the grounded substrate.

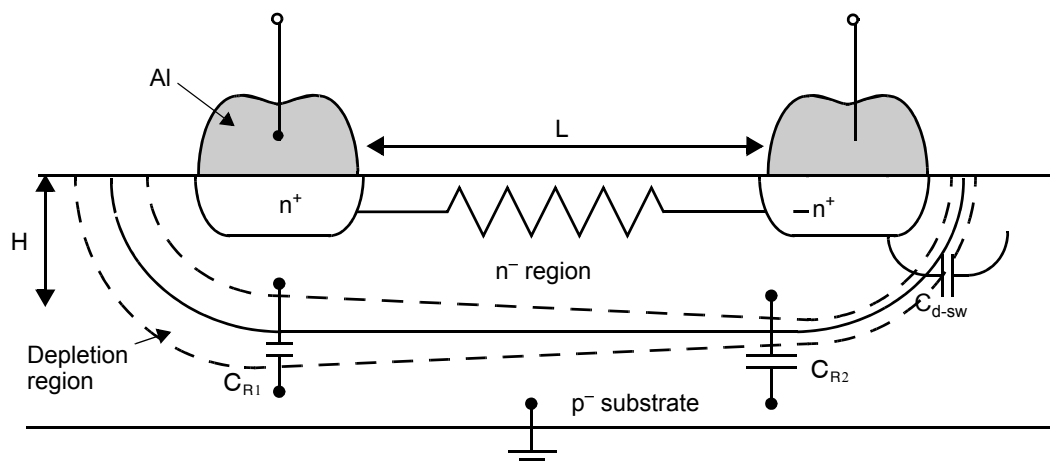


Fig. 1.35 A resistor made from a lightly doped semiconductor region with depletion region capacitances shown.

type-n resistor in a p-substrate is shown in Fig. 1.35. The resistor is isolated from the grounded substrate by a reverse-biased pn-junction. Depending on the dopant concentration, the effective sheet resistance may be in the range of 10's of Ohms with a complex temperature dependence. A detailed derivation is presented in section 1.7.4 resulting the in the following sheet resistance, assuming a uniform height (depth) resistor, H:

$$R_{\square} = \frac{1}{q\mu_n H n_n} \tag{1.134}$$

The small-signal model in Fig. 1.34 is also applicable to semiconductor resistors. However, unlike strip resistors, the parasitic capacitors are junction capacitances so their values depend on the voltages at either end of the resistor. Furthermore, the extent to which the depletion region extends into the resistor also depends upon the resistor's terminal voltages, so its effective cross-section and resistance will change with voltage. This can be problematic for analog design since capacitors and resistors whose values depend on their terminal voltages are nonlinear circuit elements and will introduce harmonic distortion when subjected

to time-varying signals. Finally, the parasitic capacitances C_{R1} and C_{R2} for semiconductor resistors are typically larger than for strip resistors because they can be located further from the grounded substrate, and because the relative permittivity of silicon ($K_s \cong 11.8$) is higher than that of the silicon dioxide insulation above the substrate ($K_{ox} \cong 3.9$).

EXAMPLE 1.20

A 4 kΩ resistor is to be formed from a single 3-μm-wide strip of type-n silicon with a dopant concentration of $N_D = 10^{23}$ atoms/m^3 that extends to a depth of $H = 2\ \mu m$ into a type-p substrate that has a dopant concentration of $N_A = 10^{22}$ atoms/m^3. Find the length of the resistor and estimate the parasitic capacitances. Assume that the resistor-substrate junction has a 3-V reverse bias, and $\mu_n = 8 \cdot 10^{-2}\ m^2/V \cdot s$ inside the resistor.

Solution

Substitution into (1.134) yields a sheet resistance of

$$R_\square = \frac{1}{(1.6 \cdot 10^{-19})(8 \cdot 10^{-2})(2 \cdot 10^{-6})(10^{23})} = 390\ \Omega/\text{sq}$$

Rearranging (1.133),

$$L = \left(\frac{R}{R_\square}\right)W = \left(\frac{4\ k\Omega}{0.39\ k\Omega}\right)3\ \mu m = 31\ \mu m$$

This lightly-doped junction has a built-in voltage lower than was calculated in Example 1.2.

$$\Phi_0 = 0.026 \times \ln\left(\frac{10^{23} \times 10^{22}}{(1.1 \times 10^{16})^2}\right) = 0.77\ V$$

The junction capacitance per unit area, assuming an abrupt one-sided junction ($N_A \ll N_D$) and using the reverse bias voltage $V_R = 3$ V, may be calculated as follows:

$$C_{j0} = \sqrt{\frac{qK_s\varepsilon_0 N_A}{2\Phi_0}} = \sqrt{\frac{1.6 \times 10^{-19} \times 11.8 \times 8.854 \times 10^{-12} \times 10^{22}}{2 \times 0.77}} = 0.33\ fF/\mu m^2$$

$$C_j = \frac{C_{j0}}{\sqrt{1 + \frac{V_R}{\Phi_0}}} = \frac{0.33}{\sqrt{1 + \frac{3}{0.77}}}\ fF/\mu m^2 = 0.15\ fF/\mu m^2$$

The resistor forms a pn-junction with an area of $WL = (3\ \mu m \cdot 31\ \mu m) = 93\ \mu m^2$ and sidewalls having an area approximately equal to the resistor perimeter times its height,

$$A_{sw} = (3\ \mu m + 31\ \mu m + 3\ \mu m + 31\ \mu m) \cdot 2\ \mu m = 136\ \mu m^2$$

Adopting the simple model of Fig. 1.35 and assuming negligible voltage drop across the resistor, the total capacitance may be divided equally between the two resistor terminals.

$$C_{R1} = C_{R2} = \frac{(WL + A_{sw})C_j}{2} = \frac{(93 + 136)0.15}{2}\ \text{fF} = 17\ \text{fF}$$

Triode MOS Resistors

There is a third method to realize resistors on an integrated circuit with which the reader is already familiar: a MOS device in triode. Pictured in Fig. 1.11, this is effectively a semiconductor resistor where the channel depth, and hence resistance, are modulated by a gate voltage. Therefore, like semiconductor resistors, triode MOS devices exhibit nonlinear resistance and parasitic capacitances. But, because the conducting channel region is much shallower than in semiconductor resistors, the resistance is even more nonlinear than a semiconductor resistor. Typically, MOS triode resistors are only used when the voltage across the resistor terminals (i.e. the drain and source of the MOS device) will remain significantly less than V_{eff}.

In spite of these shortcomings, MOS devices in triode have two unique properties that make them very useful as resistors in analog integrated-circuit design. First, the value of resistance may be tuned via a third terminal: the gate voltage. Second, relatively large values of resistance can be realized in a very compact area.

EXAMPLE 1.21

As in Example 1.19, you require a resistor of value $50\text{k}\Omega$. However, since the voltage across the resistor is never expected to exceed 200 mV, you decide it is feasible to use a n-channel MOS devices in triode to implement it. You elect to use multiple devices sized $W/L = 2\mu\text{m}/1\mu\text{m}$ to avoid short-channel effects. If $\mu_n C_{ox} = 300\mu\text{A}/\text{V}^2$, $V_{tn} = 0.3\text{V}$, and you elect to use $V_{GS} = 1\text{V}$, find the number of series devices required to implement the resistor. Compare the resulting device area to that calculated in Example 1.19 for a strip resistor.

Solution

The resistance of a device in triode is given in equation (1.104).

$$r_{ds} = \left(\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})\right)^{-1}$$

$$= \left(300\mu\text{A}/\text{V}^2\left(\frac{2\mu\text{m}}{1\mu\text{m}}\right)(1\text{V} - 0.3\text{V})\right)^{-1}$$

$$= 2.38\text{k}\Omega$$

With 21 such devices in series, the desired resistance $21 \cdot 2.38\text{k}\Omega = 50\text{k}\Omega$ is realized. This is considerably smaller than combining 20 strip resistors in series, each $5\mu\text{m} \times 1\mu\text{m}$, as in Example 1.19.

Regardless of how they are implemented, resistor values vary greatly from one integrated circuit to another and with changes in temperature. Fortunately, with some care it is possible to ensure that such variations effect all

resistors on a given integrated circuit approximately equally. Analog integrated circuits must often be designed to function properly even when the resistor values change by 10 – 40%.

1.6.2 Capacitors

pn Junction Capacitors

The capacitance provided by reverse-biased pn-junctions has already been discussed in Section 1.1. Junction capacitances may be introduced intentionally into an analog design to serve as a capacitor. They provide a relatively high value of capacitance per unit area since the depletion region can be quite thin and the relative permittivity of silicon is quite high ($K_s \cong 11.8$). Their capacitance can be tuned by adjusting the voltage across them. When used to realize a variable capacitor in this way, they are called *varactors*.

Unfortunately, several features of junction capacitances pose problems for analog design. First, although the tunability of their capacitance is useful in some applications, it also makes them nonlinear elements that distort time-varying signals. Second, the value of capacitance realized by a given size junction varies considerably with dopant concentration which is difficult to control during integrated circuit manufacturing. Finally, junction capacitors have more leakage current than other types of integrated circuit capacitors. The most common application for pn junction capacitors is as a varactor in tunable radio-frequency circuits, notably oscillators, although even there MOS capacitors are now often favoured.

MOS Capacitors

Since the gate oxide is the thinnest dielectric available in an integrated circuit, it is imminently sensible to build capacitors around it. There are many ways to do so. All comprise gate and silicon conducting "plates" separated by the gate oxide as a dielectric. All are nonlinear capacitors, whose value depends on the voltage across it. The detailed modeling of all of these is beyond the scope of this section, so we will focus on just one common structure: the PMOS transistor.

When using a PMOS transistor, one terminal is the gate and the other is the source, drain, and body all shorted together underneath, shown schematically in Fig. 1.36(*a*). In this configuration, $V_{DS} = 0$ and $V_{SG} = -V_{GS}$ is the voltage on the capacitor. Small-signal models for a NMOS transistor in this mode were covered in Section 1.2.8, but the salient points are repeated here for a PMOS device.

If $V_{SG} > |V_{tp}|$, the device enters triode and the small-signal capacitance is given by the sum of C_{gs} and C_{gd} from equation (1.105) and two overlap capacitances, C_{ov}, from equation (1.90).

$$C_{MOS(on)} = WLC_{ox} + 2WL_{ov}C_{ox} \tag{1.135}$$

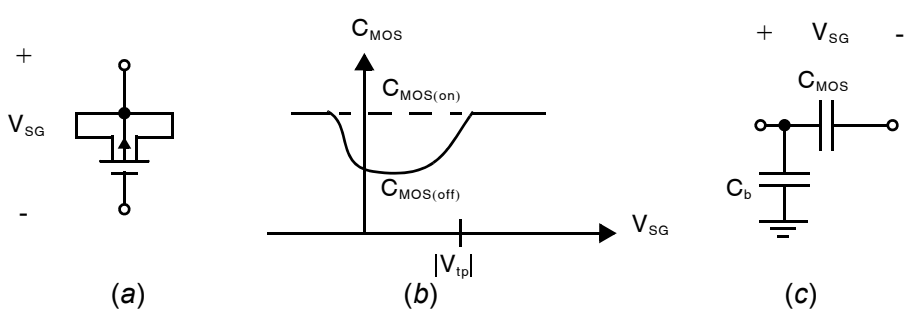


Fig. 1.36 A PMOS capacitor: (*a*) schematic symbol; (*b*) nonlinear capacitance; (*c*) small-signal model.

If $V_{SG} < |V_{tp}|$, the channel under the device turns off and the small-signal capacitance is mainly provided by only the overlap capacitances C_{gs} and C_{gd} in equation (1.110).

$$C_{MOS(off)} \cong 2WL_{ov}C_{ox} \tag{1.136}$$

Here, we have neglected C_{gb} which is assumed to be much smaller. Finally, if V_{SG} becomes significantly less than zero, the silicon immediately under the gate will eventually accumulate n-type carriers and become conducting again, thus increasing the small-signal capacitance back to a value close to (1.135). Hence, the structure's smalls-signal capacitance varies with voltage as shown in Fig. 1.36(*b*).

The PMOS body terminal is typically a n-type well in the p-type substrate, isolated from ground by a reverse-biased pn-junction. This of course introduces a junction capacitance between the body and ground, C_b, which can be quite large and whose value varies with changes in the body voltage. A simple small-signal model is shown in Fig. 1.36(*c*). For more accurate modeling, a series resistance may be included in series with C_{MOS} due to the resistance of the inverted channel or body region.

Clearly this is a highly nonlinear capacitor and should be used with care in analog design. Nevertheless, it is popular for three reasons. First, because it is relatively well-modeled by standard circuit simulators. It is, after all, a transistor. Second, the very thin gate oxide ensures a high capacitance-per-unit-area. Third, it requires no special modifications to CMOS integrated-circuit fabrication processes.

Metal-metal

To realize a purely linear capacitance on an integrated circuit, it is necessary to avoid the use of semiconductors for either plate. Instead, the many electrically-isolated layers of conductors above the silicon substrate are used.

Two different geometries used to implement metal-metal capacitors are shown in Fig. 1.37. The capacitance of the parallel-plate structure is approximately

$$C \cong \frac{\varepsilon_{ox}A}{t_{ox}} \tag{1.137}$$

where t_{ox} is the spacing between plates (on the order of 0.1 - 10 μm), ε_{ox} is the permittivity of the insulator between plates (often silicon dioxide, for which $\varepsilon_{ox} \cong 3.9\varepsilon_0$), and A is the area of the plate. Equation (1.137) neglects the fringe fields around the edges of the capacitor, but these are relatively small if the capacitor area is large. Some integrated circuits are specially engineered to provide parallel-plate capacitors with very small t_{ox} and/or large ε_{ox}, thus permitting small area A for a given value of capacitance. For example, two polysilicon (gate) layers are sometimes stacked on top of each other very closely to realize a "double-poly capacitor".[16] If no such provision is made, the normal metal layers intended for wiring may be used.

The parallel-plate structure is asymmetric since the bottom plate is in closer proximity to the silicon substrate and electrically shields the top plate from ground. Hence, the bottom plate has a larger parasitic capacitance, $C_{p1} \gg C_{p2}$. This can often be exploited in analog designs by connecting the bottom plate to a node with a constant potential with respect to ground, thus minimizing parasitics on the plate with time-varying potential. In light of this, a slightly different symbol is used in this book, as shown in Fig. 1.37(*a*), when the distinction between top and bottom plate is crucial.

The side-wall capacitive structure shown in Fig. 1.37(*b*) is becoming increasingly popular in modern integrated circuits where metal wires may actually be made thinner (when viewed from above) than their height. Hence, routing many of them immediately next to each other can provide a large total capacitance. In fact, the capacitance-per-unit-area of such a structure is often greater than parallel-plate metal-metal structures (unless special high-density parallel-plate capacitors are available). The structure is symmetric, with equal parasitics on either side of the capacitor, $C_{p1} = C_{p2}$. However, the capacitance is more difficult to estimate since it is largely provided by fringing electrical fields so appropriate computer tools are required to properly design them.

16. Although polysilicon is not a metal, it can be used to realize linear parallel-plate capacitors.

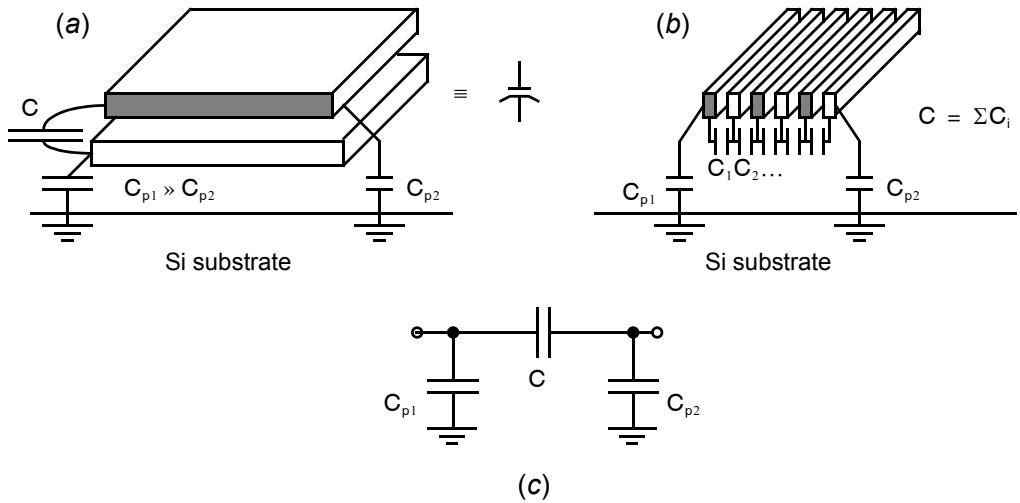


Fig. 1.37 Metal-metal capacitor geometries: (*a*) parallel plate capacitor; (*b*) side-wall capacitors; (*c*) small-signal model.

When using either of these geometries, parasitics are minimized by locating them as high above the silicon substrate as possible. Of course, this may not be possible - for example, when a double-poly capacitor is being used. Furthermore, either of these geometries, or combinations of them, may be stacked on top of each other. For example, several parallel-plates may be stacked on top of each other with alternating layers shorted to provide a metal "sandwich" capacitor. Such complex structures generally provide even higher capacitance per unit area, and require the use of computer tools to accurately estimate their value.

1.7 APPENDIX

The purpose of this appendix is to present derivations for device equations that rely heavily on device physics knowledge. Specifically, equations are derived for the exponential relationship and diffusion capacitance of diodes, and for the threshold voltage and triode relationship for MOS transistors.

1.7.1 Diode Exponential Relationship

The concentration of minority carriers in the bulk, far from the junction, is given by (1.2) and (1.4). Close to the junction, the minority-carrier concentrations are much larger. Indeed, the concentration next to the junction increases exponentially with the external voltage, V_D, that is applied in the forward direction. The concentration of holes in the n side next to the junction, p_n, is given by [Sze, 1981]

$$p_n = p_{n0} e^{V_D/V_T} = \frac{n_i^2}{N_D} e^{V_D/V_T} \tag{1.138}$$

Similarly, the concentration of electrons in the p side next to the junction is given by

$$n_p = n_{p0} e^{V_D/V_T} = \frac{n_i^2}{N_A} e^{V_D/V_T} \tag{1.139}$$

As the carriers diffuse away from the junction, their concentration exponentially decreases. The relationship for holes in the n side is

$$p_n(x) = p_n(0) e^{-x/L_p} \tag{1.140}$$

where x is the distance from the junction and L_p is a constant known as the diffusion length for holes in the n side. Similarly, for electrons in the p side we have

$$n_p(x) = n_p(0) e^{-x/L_n} \tag{1.141}$$

where L_n is a constant known as the diffusion length of electrons in the p side. Note that $p_n(0)$ and $n_p(0)$ are given by (1.138) and (1.139), respectively. Note also that the constants L_n and L_p are dependent on the doping concentrations N_A and N_D, respectively.

The current density of diffusing carriers moving away from the junction is given by the well-known diffusion equations [Sze, 1981]. For example, the current density of diffusing electrons is given by

$$J_{D\text{-}n} = -qD_n \frac{dn_p(x)}{dx} \tag{1.142}$$

where D_n is the diffusion constant of electrons in the p side of the junction. The negative sign is present because electrons have negative charge. Note that $D_n = (kT/q)\mu_n$, where μ_n is the mobility of electrons. Using (1.141), we have

$$\frac{dn_p(x)}{dx} = \frac{n_p(0)}{L_n} e^{-x/L_n} = -\frac{n_p(x)}{L_n} \tag{1.143}$$

Therefore

$$J_{D\text{-}n} = \frac{qD_n}{L_n} n_p(x) \tag{1.144}$$

Thus, the current density due to diffusion is proportional to the minority-carrier concentration. Next to the junction, all the current flow results from the diffusion of minority carriers. Further away from the junction, some of the current flow is due to diffusion and some is due to majority carriers drifting by to replace carriers that recombined with minority carriers or diffused across the junction.

Continuing, we use (1.139) and (1.144) to determine the current density next to the junction of electrons in the p side:

$$\begin{aligned} J_{D\text{-}n} &= \frac{qD_n}{L_n} n_p(0) \\ &= \frac{qD_n}{L_n} \frac{n_i^2}{N_A} e^{V_D/V_T} \end{aligned} \tag{1.145}$$

For the total current of electrons in the p side, we multiply (1.145) by the effective junction area, A_D. The total current remains constant as we move away from the junction since, in the steady state, the minority carrier concentration at any particular location remains constant with time. In other words, if the current changed as we moved away from the junction, the charge concentrations would change with time.

Using a similar derivation, we obtain the total current of holes in the n side, $I_{D\text{-}p}$, as

$$I_{D\text{-}p} = \frac{A_D q D_n n_i^2}{L_p N_D} e^{V_D/V_T} \tag{1.146}$$

where D_n is the diffusion constant of electrons in the p side of the junction, L_p is the diffusion length of holes in the n side, and N_D is the impurity concentration of donors in the n side. This current, consisting of positive carriers, flows in the direction opposite to that of the flow of minority electrons in the p side. However, since electron carriers are negatively charged, the direction of the current flow is the same. Note also that if the p side is more heavily doped than the n side, most of the carriers will be holes, whereas if the n side is more heavily doped than the p side, most of the carriers will be electrons.

The total current is the sum of the minority currents at the junction edges:

$$I_D = A_D q n_i^2 \left(\frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right) e^{V_D/V_T} \tag{1.147}$$

Equation (1.147) is often expressed as

$$I_D = I_S e^{V_D/V_T} \tag{1.148}$$

where

$$I_S = A_D q n_i^2 \left(\frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right) \tag{1.149}$$

Equation (1.148) is the well-known exponential current–voltage relationship of forward-biased diodes.

The concentrations of minority carriers near the junction and the direction of current flow are shown in Fig. 1.38.

1.7.2 Diode-Diffusion Capacitance

To find the diffusion capacitance, C_d, we first find the minority charge close to the junction, Q_d, and then differentiate it with respect to V_D. The minority charge close to the junction, Q_d, can be found by integrating either (1.140) or (1.141) over a few diffusion lengths. For example, if we assume n_{p0}, the minority electron concentration in the p side far from the junction is much less than $n_p(0)$, the minority electron concentration at the junction edge, we can use (1.141) to obtain

$$\begin{aligned} Q_n &= qA_D \int_0^\infty n_p(x)\,dx \\ &= qA_D \int_0^\infty n_p(0) e^{-X/L_n}\,dx \\ &= qA_D L_n n_p(0) \end{aligned} \tag{1.150}$$

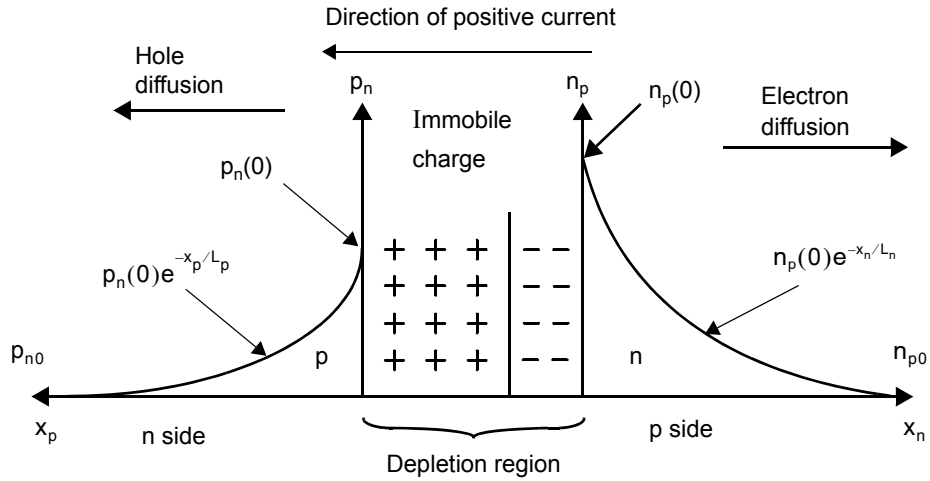


Fig. 1.38 The concentration of minority carriers and the direction of diffusing carriers near a forward-biased junction.

Using for $n_p(0)$ results in

$$Q_n = \frac{qA_D L_n n_i^2}{N_A} e^{V_D/V_T} \tag{1.151}$$

In a similar manner, we also have

$$Q_p = \frac{qA_D L_n n_i^2}{N_D} e^{V_D/V_T} \tag{1.152}$$

For a typical junction, one side will be much more heavily doped than the other side, and therefore the minority charge storage in the heavily doped side can be ignored since it will be much less than that in the lightly doped side. Assuming the n side is heavily doped, we find the total charge, Q_d, to be approximately given by Q_n, the minority charge in the p side. Thus, the small-signal diffusion capacitance, C_d, is given by

$$C_d = \frac{dQ_d}{dV_D} \cong \frac{dQ_n}{dV_D} = \frac{qA_D L_n n_i^2}{N_A V_T} e^{V_D/V_T} \tag{1.153}$$

Using (1.147) and again noting that $N_D >> N_A$, we have

$$C_d = \frac{L_n^2}{D_n} \frac{I_D}{V_T} \tag{1.154}$$

Equation (1.154) is often expressed as

$$C_d = \tau_T \frac{I_D}{V_T} \tag{1.155}$$

where τ_T is the transit time of the diode given by

$$\tau_T = \frac{L_n^2}{D_n} \tag{1.156}$$

for a single-sided diode in which the n side is more heavily doped.

1.7.3 MOS Threshold Voltage and the Body Effect

Many factors affect the gate-source voltage at which the channel becomes conductive. These factors are as follows:

1. The work-function difference between the gate material and the substrate material
2. The voltage drop between the channel and the substrate required for the channel to exist
3. The voltage drop across the thin oxide required for the depletion region, with its immobile charge, to exist
4. The voltage drop across the thin oxide due to unavoidable charge trapped in the thin oxide
5. The voltage drop across the thin oxide due to implanted charge at the surface of the silicon. The amount of implanted charge is adjusted in order to realize the desired threshold voltage.

The first factor affecting the transistor threshold voltage, V_{th}, is the built-in *Fermi potential* due to the different materials and doping concentrations used for the gate material and the substrate material. If one refers these potentials to that of intrinsic silicon [Tsividis, 1987], we have

$$\phi_{F\text{-Gate}} = \frac{kT}{q}\ln\left(\frac{N_D}{n_i}\right) \tag{1.157}$$

for a polysilicon gate with doping concentration N_D, and

$$\phi_{F\text{-Sub}} = \frac{kT}{q}\ln\left(\frac{n_i}{N_A}\right) \tag{1.158}$$

for a p substrate with doping concentration N_A. The work-function difference is then given by

$$\begin{aligned}\phi_{MS} &= \phi_{F\text{-Sub}} - \phi_{F\text{-Gate}} \\ &= \frac{kT}{q}\ln\left(\frac{N_D N_A}{n_i^2}\right)\end{aligned} \tag{1.159}$$

The next factor that determines the transistor threshold voltage is the voltage drop from the channel to the substrate, which is required for the channel to exist. The question of exactly when the channel exists does not have a precise answer. Rather, the channel is said to exist when the concentration of electron carriers in the channel is equal to the concentration of holes in the substrate. At this gate voltage, the channel is said to be *inverted.* As the gate voltage changes from a low value to the value at which the channel becomes inverted, the voltage drop in the silicon also changes, as does the voltage drop in the depletion region between the channel and the bulk. After the channel becomes inverted, any additional increase in gate voltage is closely equal to the increase in voltage drop across the thin oxide. In other words, after channel inversion, gate voltage variations have little effect on the voltage drop in the silicon or the depletion region between the channel and the substrate.

The electron concentration in the channel is equal to the hole concentration in the substrate when the voltage drop from the channel to the substrate is equal to two times the difference between the Fermi potential of the substrate and intrinsic silicon, ϕ_F, where

$$\phi_F = -\frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) \tag{1.160}$$

Equation (1.200) is a factor in several equations used in modelling MOS transistors. For typical processes, ϕ_F can usually be approximated as 0.35 V for typical doping levels at room temperature.

The third factor that affects the threshold voltage is due to the immobile negative charge in the depletion region left behind after the p mobile carriers are repelled. This effect gives rise to a voltage drop across the thin oxide of $-Q_B/C_{ox}$, where

$$Q_B = -qN_A x_d \tag{1.161}$$

and x_d is the width of the depletion region. Since

$$x_d = \sqrt{\frac{2K_s\varepsilon_0|2\phi_F|}{qN_A}} \tag{1.162}$$

we have

$$Q_B = -\sqrt{2qN_AK_s\varepsilon_0|2\phi_F|} \tag{1.163}$$

The fourth factor that determines V_{tn} is due to the unavoidable charge trapped in the thin oxide. Typical values for the effective ion density of this charge, N_{ox}, might be 2×10^{14} to 10^{15} ions/m^3. These ions are almost always positive. This effect gives rise to a voltage drop across the thin oxide, V_{ox}, given by

$$V_{ox} = \frac{-Q_{ox}}{C_{ox}} = \frac{-qN_{ox}}{C_{ox}} \tag{1.164}$$

The *native transistor threshold voltage* is the threshold voltage that would occur naturally if one did not include a special ion implant used to adjust the threshold voltage. This value is given by

$$V_{t\text{-native}} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} \tag{1.165}$$

A typical native threshold value might be around –0.1 V. It should be noted that transistors that have native transistor threshold voltages are becoming more important in analog circuit design where they are used in transmission gates or in source-follower buffers.

The fifth factor that affects threshold voltage is a charge implanted in the silicon under the gate to change the threshold voltage from that given by (1.165) to the desired value, which might be between 0.2 to 0.7 V for an n-channel transistor.

For the case in which the source-to-substrate voltage is increased, the effective threshold voltage is increased. This is known as the *body effect*. The body effect occurs because, as the source-bulk voltage, V_{SB}, becomes larger, the depletion region between the channel and the substrate becomes wider, and therefore more immobile negative charge becomes uncovered. This increase in charge changes the third factor in determining the transistor threshold voltage. Specifically, instead of using (1.163) to determine Q_B, one should now use

$$Q_B = -\sqrt{2qN_AK_s\varepsilon_0(V_{SB} + |2\phi_F|)} \tag{1.166}$$

If the threshold voltage when $V_{SB} = 0$ is denoted V_{tn0}, then, using (1.165) and (1.166), one can show that

$$\begin{aligned} V_{tn} &= V_{tn0} + \Delta V_{tn} \\ &= V_{tn0} + \frac{\sqrt{2qN_AK_s\varepsilon_0}}{C_{ox}}\left[\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right] \\ &= V_{tn0} + \gamma\left(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right) \end{aligned} \tag{1.167}$$

where

$$\gamma = \frac{\sqrt{2qN_A K_s \varepsilon_0}}{C_{ox}} \tag{1.168}$$

The factor γ is often called the *body-effect constant.*

1.7.4 MOS Triode Relationship

The current flow in a MOS transistor is due to *drift* current rather than diffusion current. This type of current flow is the same mechanism that determines the current in a resistor. The current density, J, is proportional to the electrical field, E, where the constant of proportionality, σ, is called the *electrical permittivity.* Thus,

$$J = \sigma E \tag{1.169}$$

This constant for an n-type material is given by

$$\sigma = qn_n\mu_n \tag{1.170}$$

where n is the concentration per unit volume of negative carriers and μ_n is the mobility of electrons. Thus, the current density is given by

$$J = qn_n\mu_n E \tag{1.171}$$

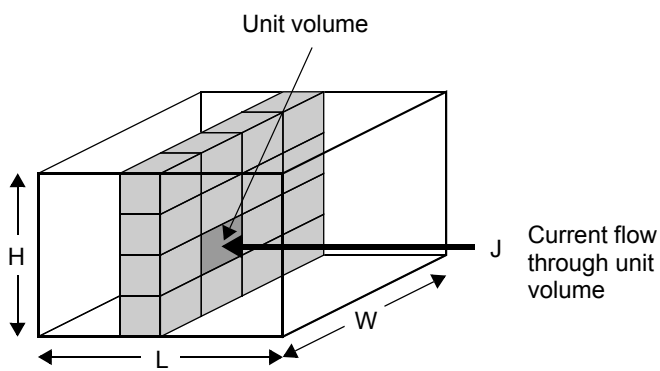


Fig. 1.39 Current flowing through a unit volume.

Next, consider the current flow through the volume shown in Fig. 1.39, where the volume has height H and width W. The current is flowing perpendicular to the plane $H \times W$ down the length of the volume, L. The current, I, everywhere along the length of the volume is given by

$$I = JWH \tag{1.172}$$

The voltage drop along the length of the volume in the direction of L for a distance dx is denoted dV and is given by

$$dV = E(x)dx \tag{1.173}$$

Combining (1.171), (1.172), and (1.173), we obtain

$$q\mu_n WHn_n(x)\, dV = I\, dx \tag{1.174}$$

where the carrier density $n_n(x)$ is now assumed to change along the length L and is therefore a function of x.

As an aside, we examine the case of a resistor where $n_n(x)$ is usually constant. A resistor of length L would therefore have a current given by

$$I = \frac{q\mu_n WHn_n}{L}\Delta V \tag{1.175}$$

Thus, the resistance is given by

$$R = \frac{L}{q\mu_n WHn_n} \tag{1.176}$$

Often this resistance is presented in a relative manner, in which the length and width are removed (since they can be design parameters) but the height remains included. In this case, the resulting expression is commonly referred to as the *resistance per square* and designated as $R_\square$ where

$$R_\square = \frac{1}{q\mu_n H n_n} \tag{1.177}$$

The total resistance is then given by equation (1.133), repeated here for convenience.

$$R_{total} = R_\square \frac{L}{W} \tag{1.178}$$

In the case of a MOS transistor, the charge density is not constant down the channel. If, instead of the carrier density per unit volume, one expresses $n_n(x)$ as a function of charge density per square area from the top looking down, we have

$$Q_n(x) = qHn_n(x) \tag{1.179}$$

Substituting (1.179) into (1.174) results in

$$\mu_n W Q_n(x)\, dV = I\, dx \tag{1.180}$$

Equation (1.180) applies to drift current through any structure that has varying charge density in the direction of the current flow. It can also be applied to a MOS transistor in the triode region to derive its I-V relationship. It should be noted here that in this derivation, it is assumed the source voltage is the same as the substrate voltage.

Since the transistor is in the triode region, we have $V_{DG} < -V_{tn}$. This requirement is equivalent to $V_{DS} < V_{GS} - V_{tn} = V_{eff}$. It is assumed that the effective channel length is L. Assuming the voltage in the channel at distance x from the source is given by $V_{ch}(x)$, from Fig. 1.40, we have

$$Q_n(x) = C_{ox}[V_{GS} - V_{ch}(x) - V_{tn}] \tag{1.181}$$

Substituting (1.181) into (1.180) results in

$$\mu_n W C_{ox}[V_{GS} - V_{ch}(x) - V_{tn}]dV_{ch} = I_D dx \tag{1.182}$$

Integrating both sides of (1.182), and noting that the total voltage along the channel of length L is V_{DS}, we obtain

$$\int_0^{V_{DS}} \mu_n W C_{ox}[V_{GS} - V_{ch}(x) - V_{tn}]dV_{ch} = \int_0^L I_D dx \tag{1.183}$$

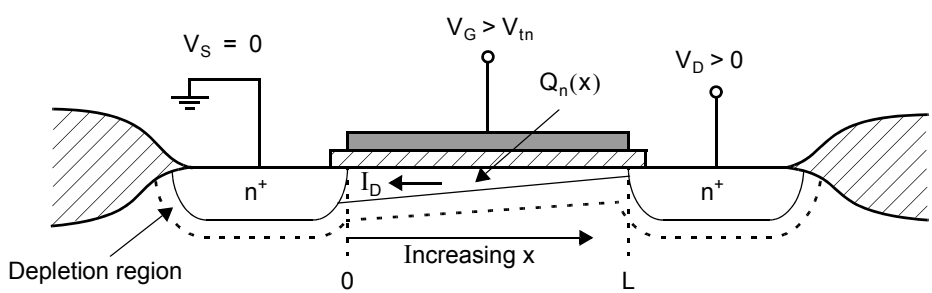


Fig. 1.40 The transistor definitions used in developing the transistor's I-V relationship.

which results in

$$\mu_n W C_{ox}\left[(V_{GS}-V_{tn})V_{DS}-\frac{V_{DS}^2}{2}\right] = I_D L \tag{1.184}$$

Thus, solving for I_D results in the well-known triode relationship for a MOS transistor:

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS}-V_{tn})V_{DS}-\frac{V_{DS}^2}{2}\right] \tag{1.185}$$

It should be noted that taking into account the body effect along the channel, the triode model of (1.225) is modified to

$$I_D = \mu C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS}-V_{tn})V_{DS}-\alpha\frac{V_{DS}^2}{2}\right] \tag{1.186}$$

where $\alpha \cong 1.7$ [Tsividis, 1987].

1.8 KEY POINTS

- The charge–voltage relationship of a reverse-biased pn junction is modeled by a nonlinear depletion capacitance. [p. 6]
- The source terminal of an n-channel transistor is defined as whichever of the two terminals has a lower voltage. For a p-channel transistor, the source would be the terminal with the higher voltage. [p. 14]
- The relationship between drain-source voltage and drain current of a MOS device is approximately linear when $V_{DS} << V_{eff}$. [p. 19]
- For $V_{DS} > V_{eff}$, a MOS device exhibits a square-law current–voltage relationship. [p. 21]
- The body effect is the influence of the body potential on the channel, modelled as an increase in the threshold voltage, V_{tn}, with increasing source-to-body reverse-bias. [p. 24]
- The square-root relationship for transconductance is useful for circuit analysis when device sizes are fixed. However, the simpler expression $g_m = 2I_D/V_{eff}$ is useful during initial circuit design when transistor sizes are yet to be determined. [p. 26]
- Small signal r_{ds} is proportional to L/I_D. [p. 28]
- In a MOSFET, the largest parasitic capacitance is C_{gs}, proportional to gate area WL and via C_{ox} inversely proportional to oxide thickness. [p. 31]
- The gate-drain capacitance C_{gd}, also known as the Miller capacitance, is due to physical overlap of the gate and drain regions as well as fringing fields. It is especially important when there is a large voltage gain between gate and drain. [p. 32]
- For operation with high gain, transistors should have long gate lengths and be biased with low V_{eff}. For high speed, the opposite is desirable: small L and high V_{eff}. [p. 38]
- In subthreshold operation, also called weak inversion, transistors obey an exponential voltage–current relationship instead of a square-law. Hence, a small but finite current flows even when $V_{GS} = 0$. [p. 42]
- For a fixed drain current, the small-signal transconductance of a MOS device is maximized in subthreshold operation at a value of $g_m = qI_D/nkT$. [p. 43]
- For large values of V_{eff}, transistors have a sub-square-law voltage current relationship, tending towards a linear relationship for very large values of V_{eff}. [p. 46]

1.9 REFERENCES

R. Geiger, P. Allen, and N. Strader. *VLSI: Design Techniques for Analog and Digital Circuits*. McGraw-Hill, New York, 1990.

P. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits,* 5th. ed. John Wiley & Sons, New York, 2009.

D. Hodges and H. Jackson. *Analysis and Design of Digital Integrated Circuits,* 2nd ed. McGraw-Hill, New York, 1988.

D. Roulston. *Semiconductor Devices.* McGraw-Hill, New York, 1990.

C. G. Sodini, P.-K. Ko, and J. L. Moll, "The effect of High Fields on MOS Device and Circuit Performance," *IEEE Trans. Electron Devices*, Vol. 31, pp. 1386-1393, October 1984.

S. M. Sze. *Physics of Semiconductor Devices.* Wiley Interscience, New York, 1981.

Y. Tsividis. *Operation and Modeling of the MOS Transistor.* McGraw-Hill, New York, 1987.

S. Wolf. *Silicon Processing for the VLSI Era—Volume 3: The Submicron MOSFET.* Lattice Press, Sunset Beach, California, 1995.

1.10 PROBLEMS

1.10.1 Section 1.1: Semiconductors and pn Junctions

1.1 Estimate the hole and electron concentrations in silicon doped with arsenic at a concentration of 10^{25} atoms/m^3 at a temperature 22 °C above room temperature. Is the resulting material n type or p type?

1.2 For the pn junction of Example 1.2, estimate the new built-in potential, Φ_0, when the temperature is increased 11 °C above room temperature.

1.3 Calculate the amount of charge per $(\mu m)^2$ in each of the n and p regions of the pn junction of Example 1.2 for a 3-V reverse-bias voltage. How much charge on each side would be present in a 10 μm × 10 μm diode?

1.4 A silicon diode has C_{j0} = 15 fF. It is biased by a 43-kΩ resistor connected between the cathode of the diode and the input signal, as shown in Fig. P1.4. Initially, the input is 3 V, and then at time 0 it changes to 0 V. Estimate the time it takes for the output voltage to change from 3 V to 0.9 V (i.e., the $\Delta t_{-70\%}$ time). Repeat for an input voltage change from 0 V to 3 V and an output voltage change from 0 V to 2.1 V. Compare your answers to those obtained using a SPICE simulation.

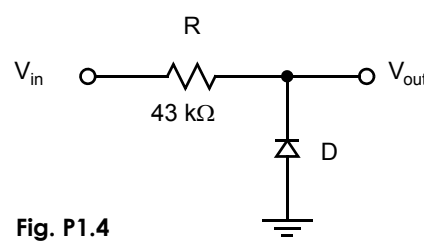


Fig. P1.4

1.5 Sketch a plot of the electric field strength versus depth for the pn junction of Example 1.3.

1.6 At what reverse-bias voltage will the pn junction of Example 1.3 breakdown? Assume that when the electric field exceeds 3×10^7 V/m, avalanche breakdown occurs.

1.7 A pn junction with $N_A \ll N_D$ and an area 40 μm^2 is observed to have a capacitance of 30 fF while under 1-V reverse bias. Estimate the dopant concentration on the p side, N_A.

1.10.2 Section 1.2: MOS Transistors and Section 1.3: Device Model Summary

1.8 Verify that when $V_{DS} = V_{eff}$ is used in the triode equation for a MOS transistor, the current equals that of the active region equation given in (1.63).

1.9 Find I_D for an n-channel transistor having doping concentrations of $N_D = 10^{26}$ atoms/m^3 and $N_A = 10^{23}$ atoms/m^3 with $W = 5\ \mu m$, $L = 0.5\ \mu m$, $V_{GS} = 1$ V, and $V_{DS} = V_{eff}$. Assuming λ remains constant, estimate the new value of I_D if V_{DS} is increased by 0.3 V.

1.10 A MOS transistor in the active region is measured to have a drain current of 20 μA when $V_{DS} = V_{eff}$. When V_{DS} is increased by 0.5 V, I_D increases to 23 μA. Estimate the output impedance, r_{ds}, and the output impedance constant, λ.

1.11 Derive the low-frequency model parameters for an n-channel transistor having doping concentrations of $N_D = 10^{26}$ atoms/m^3 and $N_A = 10^{23}$ atoms/m^3 with $W = 8\ \mu m$, $L = 0.6\ \mu m$, $V_{GS} = 0.9$ V, and $V_{DS} = V_{eff}$. Assume that $V_{SB} = 1.0$ V.

1.12 Find the capacitances C_{gs}, C_{gd}, C_{db}, and C_{sb} for an active n-channel transistor having doping concentrations of $N_D = 10^{26}$ atoms/m^3 and $N_A = 10^{23}$ atoms/m^3 with $W = 15\ \mu m$, $L = 0.5\ \mu m$. Assume that the source and drain junctions extend 1 μm beyond the gate, resulting in source and drain areas being $A_s = A_d = 15\ \mu m^2$ and the perimeter of each being $P_s = P_d = 32\ \mu m$.

1.13 Consider the circuit shown in Fig. P1.13, where V_{in} is a dc signal of 1 V. Taking into account only the channel charge storage, determine the final value of V_{out} when the transistor is turned off, assuming half the channel charge goes to C_L. You may use the parameters for the 0.18-μm CMOS processes in Table 1.5.

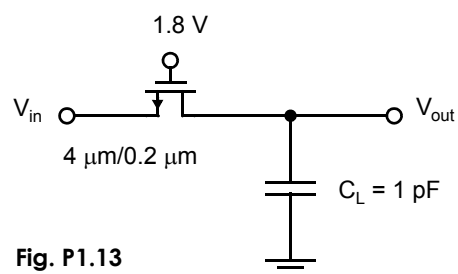


Fig. P1.13

1.14 For the same circuit as in Problem 1.13, the input voltage has a step voltage change at time 0 from 0.2 V to 0.4 V (the gate voltage remains at 1.8 V). Find its 99 percent settling time (the time it takes to settle to within 1 percent of the total voltage change). You may ignore the body effect and all capacitances except C_L. Repeat the question for V_{in} changing from 0.6 V to 0.8 V.

1.15 Repeat Problem 1.14, but now take into account the body effect on V_{tn}. Assume $N_A = 10^{23}$ atoms/m^3.

1.16 For each of the CMOS processes tabulated in Table 1.5, how many charge carriers q are required to elevate the gate voltage of a triode MOSFET by 100 mV? Assume that the gate length L is the minimum permitted in each technology and $W/L = 20$.

1.17 Using the device parameters for the 0.35-μm technology NMOS device in Table 1.5 and $L = 0.4\ \mu m$, select the device width and V_{GS} required to bias a transistor with an intrinsic gain of $A_i = 10$ and transconductance $g_m = 0.5$ mA/V. What dc current consumption is required?

1.18 Repeat Problem 1.18 using the device parameters for the 0.18-μm technology PMOS device in Table 1.5 and $L = 0.2\ \mu m$.

1.19 A NMOS transistor is to be biased with $V_{eff} = 250$ mV. Size the transistor using the device parameters for the 0.35-μm CMOS process in Table 1.5 to provide a drain current of 0.2 mA and an output resistance of $r_{ds} = 20$ kΩ. Repeat for a PMOS transistor. Repeat for the 0.18-μm CMOS process.

1.20 A NMOS transistor is to be biased with $I_D = 0.35$ mA and an intrinsic gain of $A_i = 35$ in the 0.35-μm CMOS process in Table 1.5. Find the required transistor dimentions, W and L.

1.21 A NMOS transistor is to be biased with $I_D = 0.25$ mA and transconductance of $g_m = 2.2$ mA/V in the 0.18-μm CMOS process in Table 1.5. Find the required transistor dimentions, W and L.

1.22 Consider the circuit in Fig. P1.13. Size the NMOS transistor so that, with $v_{in} = v_{out} = 0.3$ V, the circuit has a bandwidth of 250 MHz. Use the device parameters for the 0.35-μm technology in Table 1.5. Assume $L = 0.35$ μm and the threshold voltage is increased by 70 mV due to the body effect. What is the resulting gate capacitance?

1.23 Repeat Problem 1.22, this time using the device parameters for the 0.18-μm technology in Table 1.5 with $L = 0.18$ μm.

1.10.3 Section 1.4: Advanced MOS Modelling and Section 1.5: SPICE Modelling Parameters

1.24 If transistor drain current is plotted versus gate-source voltage on a log-log scale, what is the slope of the curve in strong inversion? What is the slope for very large values of V_{GS}? What is the slope of the curve at $V_{eff} = 1/\theta$?

1.25 Make a qualitative sketch of transistor intrinsic gain, A_i, versus V_{eff} for:

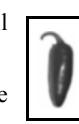
a. constant device width, W

b. constant drain current, I_D

In each case, what is the relationship between A_i and V_{eff} in weak-inversion, active mode, and under mobility degradation?

1.26 Derive expressions for transistor f_T in terms of fundamental device constants and operating point while operating in subthreshold and under mobility degradation. How do these compare with the expressions for strong inversion?

1.27 Using the Spice models from the text web site, perform simulations to extract approximate values for all of the transistor parameters listed in Table 1.5. Compare your results to the values in the table.



1.28 Two transistors with the same device paramters are biased with the same terminal voltages and have the same gate length L.

a. One of the transistors has a drain current $I_D = 0.2$ mA and a width of $W = 3$ μm. Find the width of the other transistor so that it has a transconductance $I_D = 2$ mA.

b. One of the transistors has a transconductance $g_m = 0.4$ mA/V and a width of $W = 2$ μm. Find the width of the other transistor so that it has a transconductance $g_m = 2$ mA/V.

c. One of the transistors has a small-signal $r_{ds} = 1$ kΩ and a width of $W = 30$ μm. Find the width of the other transistor so that it has a small-signal $r_{ds} = 10$ kΩ.

1.10.4 Section 1.6: Passive Devices

1.29 Assume a strip resistor has a sheet resistance of 1 kΩ/sq. and a total capacitance to ground of approximately 0.4 fF/μm^2. What is the RC time constant formed by a 4 kΩ resistor with one end grounded assuming the resistor is 1 μm wide? What if it is only 0.4 μm wide? How does this compare with the time constant obtained using the n-well resistor in Example 1.20?

1.30 A reverse-biased pn junction with $N_D = 10^{26}$ atoms/m^3 and $N_A = 10^{23}$ atoms/m^3 is to be used as a varactor. What junction area and voltage range is needed to provide a capacitance tunable from 0.2 pF to 0.3 pF?

1.31 In a particular process, a metal-metal parallel plate capacitor can be realized with a density of 7 fF/μm^2 and a metal-metal sidewall capacitor can provide a density of 10 fF/μm^2. What would be the area of a 1 pF capacitor realized using each approach? Compare this with the area of a 1 pF MOS capacitor in a 45-nm CMOS processes using the parameters in Table 1.5.

1.32 Estimate the maximum percentage change in capacitance achievable using a MOS varactor in the 0.18-μm CMOS processes in Table 1.5.

CHAPTER

2 Processing and Layout

This chapter describes the steps and processes used in realizing modern integrated circuits with emphasis on CMOS processing. After processing is presented, circuit layout is covered. Layout is the design portion of integrated-circuit manufacturing, in which the geometry of circuit elements and wiring connections is defined. This process leads to the development of photographic masks used in manufacturing an integrated circuit. The concepts of design rules and their relationship to integrated circuits are emphasized. Next, circuit layout is related to the transistor models. Here, it is shown that once the layout is completed, the values of certain elements in the transistor models can be determined. This knowledge is necessary for accurate computer simulation of integrated circuits. It is also shown that, by using typical design rules, one can make reasonable assumptions to approximate transistor parasitic components before the layout has been done. Variability in device parameters is unavoidable and particularly problematic for analog circuits. These variations are modeled and their impact on analog circuits analyzed. Finally, analog layout issues are then discussed, including matching and noise considerations.

2.1 CMOS PROCESSING

In this section, the basic steps involved in processing a CMOS integrated circuit are presented. For illustrative purposes, we describe here an example n-well process with a p substrate and two layers of metal. Many of the possible variations during processing are also described, but we focus primarily on the basics which any analog designer should know. The processing of nanometer feature sizes can require many additional steps.

2.1.1 The Silicon Wafer

The first step in realizing an integrated circuit is to fabricate a defect-free, single-crystalline, lightly doped silicon wafer. To realize such a wafer, one starts by creating metallurgical-grade silicon through the use of a high-temperature chemical process in an electrode-arc furnace. Although metallurgical-grade silicon is about 98 percent pure, it has far too many impurities for use in realizing integrated circuits. Next, a silicon-containing gas is formed and then reduced. Pure silicon is precipitated onto thin rods of single-crystalline silicon. This deposited electronic-grade silicon is very pure but, unfortunately, it is also polycrystalline. To obtain single-crystalline silicon, the silicon is melted once again and allowed to cool. As it cools, a single-crystalline *ingot* is slowly pulled and turned from the molten silicon using the *Czochralski method.* The Czochralski method starts with a seed of single crystal silicon, and the pull rate and speed of rotation determine the diameter of the crystalline rod or ingot. Typical diameters are 10 to 30 cm (i.e., 4 to 12 inches) with lengths usually longer than 1 meter. Producing a silicon ingot can take several days.

Key Point: *The first step in realizing an integrated circuit is to produce a single-crystalline silicon wafer from 10 to 30 cm in diameter and roughly 1 mm thick. The silicon is either lightly doped, or heavily doped with a thin lightly-doped epitaxial layer on top in which the transistors are made.*

Normally, heavily doped silicon is added to the *melt* before the single-crystalline ingot is pulled. After the doped silicon diffuses through the molten silicon, a more lightly doped silicon ingot results. In our example process, boron impurities would be added to produce a p-type ingot. The ingot is cut into wafers using a large diamond saw. A typical wafer might have a thickness of about 1 mm. After the ingot is sawed into wafers, each wafer is polished with Al_2O_3, chemically etched to remove mechanically damaged material, and then fine-polished again with SiO_2 particles in an aqueous solution of $NaOH$. Very often, the company that produces the silicon wafers is not the same company that eventually patterns them into monolithic circuits.

There are two methods for establishing the background doping level of the surface silicon in which all of the transistors will be made. One is to simply control the boron impurity levels in the ingot to provide a p^- wafer concentration of around $N_A \cong 2 \times 10^{21}$ donor / m^3. Such a doping level would give a resistivity of 10 to 20 $\Omega \cdot$cm. The other method is to begin with a very heavily doped p^{++} wafer with a low resistivity of around 0.01 $\Omega \cdot$cm. Then, upon the surface of the p^{++} wafer, a layer of p^- silicon is grown with a higher resistivity of 5 to 20 $\Omega \cdot$cm. All of the devices are fabricated within this top *epitaxial* layer, which may be from 2 to 20 μm thick. The use of an epitaxial layer provides more precise control over dopant concentrations while the p^{++} substrate underneath provides a low-resistivity ground plane under the circuit helping to prevent latchup, described in Section 2.2.4. In either case, transistors are fabricated in p^- silicon near the surface of the wafer.

2.1.2 Photolithography and Well Definition

Photolithography is a technique in which selected portions of a silicon wafer can be *masked* so that some type of processing step can be applied to the remaining areas. Although photolithography is used throughout the manufacture of an integrated circuit, here we describe this photographic process in the context of preparing the wafer for defining the well regions.[1]

Selective coverage for well definition is performed as follows. First, a glass mask, M_1, is created, which defines where the well regions will be located. The glass mask is created by covering the mask in photographic materials and exposing it to an electron beam, or *e beam,* in the regions corresponding to the well locations. Such exposure results in the well regions on the glass mask turning opaque, or dark. As a result, the glass mask can be thought of as a negative of one layer of the microcircuit. In a typical microcircuit process, ten to twenty different masks might be required. The cost for these masks varies considerably depending upon the minimum feature sizes to be patterned on the microcircuit. For example, currently a set of masks for a 0.35-μm CMOS process might cost roughly $30,000, whereas the cost of a mask set for the most advanced modern processes approaches $1,000,000. Because a high degree of precision is required in manufacturing these masks, often (but not always) a company other than the integrated circuit processing company makes the masks. The creation of the opaque regions of the mask by the electron beam is controlled by a computer dependent on the contents of a database. The database required for the e beam is derived from the layout database produced by the designer, using a computer-aided design (CAD) software program.

The first step in masking the surface of the wafer is to thermally grow a thin layer of silicon-dioxide (SiO_2) to protect the surface of the microcircuit. Details of this step are discussed later. On top of the SiO_2, a negative *photoresist*, PR_1, is evenly applied to a thickness of around 1 μm while spinning the microcircuit. Photoresist is a light-sensitive polymer (similar to latex). In the next step, the mask, M_1, is placed in close proximity to the wafer, and ultraviolet light is projected through the mask onto the photoresist. Wherever the light strikes, the polymers cross-link, or *polymerize*. This change makes these regions insoluble to an organic solvent. This step is shown in Fig. 2.1

1. Wells are doped regions that will contain one of the two types of transistors realized in a CMOS process. For example, wells that are n type contain p-channel transistors.

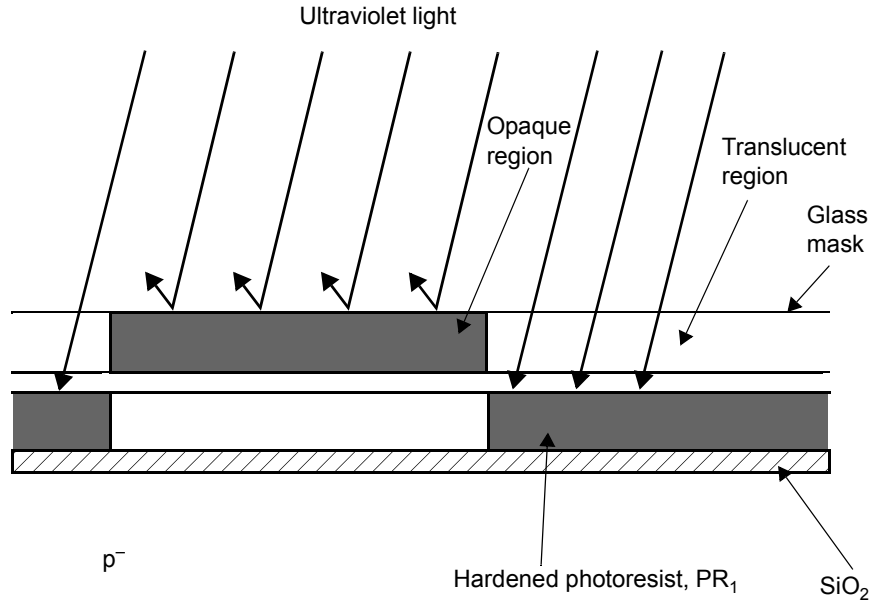


Fig. 2.1 Se-lectively hardening a region of photoresist using a glass mask.

The regions where the mask was opaque (i.e., the well regions) are not exposed. The photoresist is removed in these areas using an organic solvent. Next, the remaining photoresist is baked to harden it. After the photoresist in the well regions is removed, the uncovered SiO_2 may also be removed using an acid etch. (However, in some processes where this layer is very thin, it may not be removed.) In the next step, the dopants needed to form the well are introduced into the silicon using either diffusion or ion implantation (directly through the thin oxide, in cases where it has not been removed).

The procedure just described involves a *negative photor esist,* where the exposed photoresist remains after the masking. There are also *positive photoresists,* in which the exposed photoresist is dissolved by the organic solvents. In this case, the photoresist remains where the mask was opaque. By using both positive and negative resists, a single mask can sometimes be used for two steps—first, to protect one region and implant the complementary region and second, to protect the complementary region and implant the original region.

Key Point: *Most features on integrated circuits are patterned using photolithography whereby light is passed through a mask to cast patterns onto the underlying silicon wafer, ultimately defining the circuit's physical features such as transistor sizes and wiring.*

The feature sizes that may be patterned using photolithography are influenced by the wavelength of light used. When the integrated circuit features are smaller than the wavelength of light (currently 193-nm ultraviolet light is used), the wave nature of light results in patterns on the photoresist that do not precisely match those of the mask. Fortunately, these effects can be partially compensated for by modifying the mask pattern so that the resulting geometries more closely match those intended by the designer. This technique is referred to as "optical proximity correction" and is a common practice to realize feature sizes below 100 nm. Further improvements have been made by immersing the photolithography in a liquid bath. Doing so changes the path of light resulting in improved resolution and improved tolerance to unevenness of the substrate surface. Efforts are ongoing to reduce the minimum feature sizes that may be patterned by using shorter wavelengths for photolithography (extreme ultraviolet light or even X-rays).

2.1.3 Diffusion and Ion Implantation

After the photoresist over the well region has been removed, the next step is to introduce dopants through the opening where the well region will be located. There are two approaches for introducing these dopants: diffusion and ion implantation.

In both implantation methods, usually the SiO_2 in the well region will first be removed using an acid etch. Next, the remaining hardened photoresist is stripped using acetone. This leaves SiO_2 that was protected by the hardened photoresist to mask all of the nonwell (i.e., substrate) regions.

In diffusion implantation, the wafers are placed in a quartz tube inside a heated furnace. A gas containing the dopant is introduced into the tube. In the case of forming an n well, the dopant in the gas would probably be phosphorus. Arsenic could also be used, but it takes a much longer time to diffuse. The high temperature of the diffusion furnace, typically 900 to 1 100 °C, causes the dopants to diffuse into the silicon both vertically and horizontally. The dopant concentration will be greatest at the surface and will decrease following a *Gaussian* profile further into the silicon. If a p well had been desired, then boron would have been used as the dopant. The resulting cross section, after diffusing the n well, is shown in Fig. 2.2.

An alternative technique for introducing dopants into the silicon wafer is ion implantation. This technique has now largely replaced diffusion because it allows more independent control over the dopant concentration and the thickness of the doped region. In ion implantation, dopants are introduced as ions into the wafer, as shown in the functional representation of an ion implanter in Fig. 2.3 The ions are generated by bombarding a gas with electrons from an arc-discharge or cold-cathode source. The ions are then focused and sent through a mass separator. This mass separator bends the ion beam and sends it through a narrow slit. Since only ions of a specific mass pass through the slit, the beam is purified. Next, the beam is again focused and accelerated to between 10 keV and 1 MeV. The ion current might range from 10 μA to 2 mA. The deflection plates sweep the beam across the wafer (which is often rotated at the same time) and the acceleration potential controls how deeply the ions are implanted. The beam current and time of implantation determine the amount of dosage. Thus, depth and dosage are controlled independently. Two problems that occur with ion implantation are lattice damage and a narrow doping profile. The lattice damage is due to nuclear collisions that result in the displacement of substrate atoms. The narrow profile results in a heavy concentration over a narrow distance, as is shown in Fig. 2.4. For example, arsenic ions with an acceleration voltage of 100 keV might penetrate approximately 0.06 μm into the silicon, with the majority of ions being at 0.06 μm ± 0.02 μm. Both of these problems are largely solved by annealing.

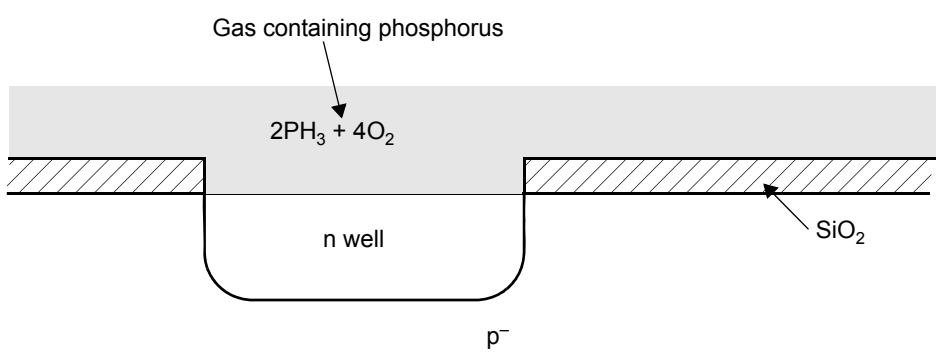


Fig. 2.2 Forming an n well by diffusing phosphorus from a gas into the silicon, through the opening in the SiO_2.

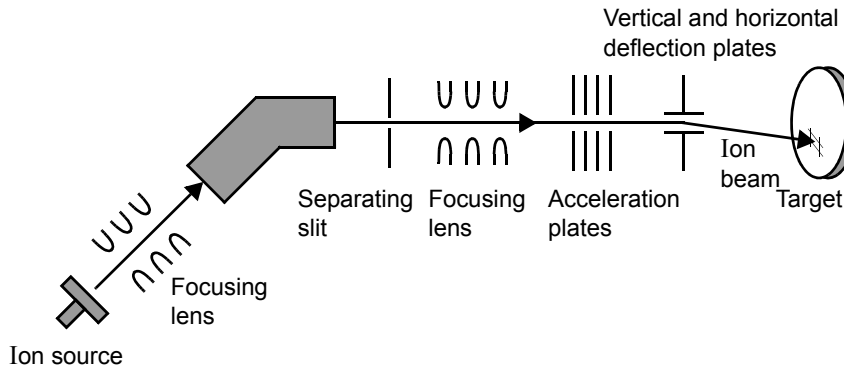


Fig. 2.3 An ion-implantation system.

Annealing is a step in which the wafer is heated to about 1000 °C, perhaps for 15 to 30 minutes, and then allowed to cool slowly. This heating stage thermally vibrates the atoms, which allows the bonds to reform. Annealing also broadens the dopant concentration profile, making the doping levels more uniform, as shown in Fig. 2.4. It should be noted that annealing is performed only once during processing, after all the implantation steps have been performed but before any metal layers have been created.[2]

For n-type dopants, arsenic is used for shallow implantations, such as the source or drain junctions. Phosphorus might be used for the well. Boron is always used to form the p regions.

Although more expensive, ion implantation has been largely replacing diffusion for forming n and p regions because of its greater control over doping levels. Another important advantage of ion implantation is the much smaller sideways diffusion, which allows devices to be more closely spaced and, more importantly for MOS transistors, minimizes the overlap between the gate-source or gate-drain regions.

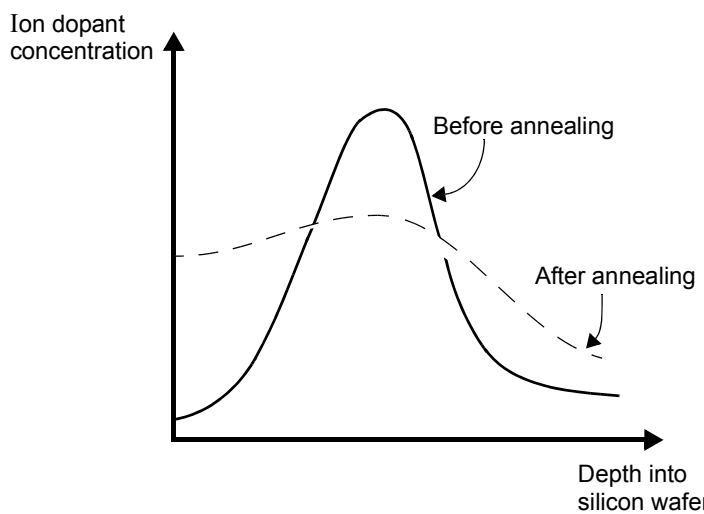


Fig. 2.4 Dopant profiles after ion implantation both before and after annealing.

2. If annealing were done after deposition of a metal layer, the metal would melt.

In some modern processes, both p and n wells are created in which NMOS and PMOS devices will be fabricated, respectively. This is referred to as a *twin-well* process.

2.1.4 Chemical Vapor Deposition and Defining the Active Regions

The next few steps use the mask M_2 to define where the transistors will be located and to make isolation structures between them. A thin layer of thermal SiO_2 is first grown everywhere to protect the surface of the silicon lattice. Next, Si_3N_4 is deposited everywhere during a gas-phase reaction in which energy is supplied by heat (at about 850 °C). This process is called *chemical vapor deposition,* or CVD. After this step, the photoresist PR_2 is deposited, exposed through the mask, M_2, dissolved, and hardened. Often, this step will be done using positive photoresist such that, wherever the mask M_2 is not opaque, the photoresist will be softened. In other words, the photoresist is left intact after the organic dissolution step, under the opaque regions of the mask. These become the active device regions where the transistors will eventually be located, also sometimes referred to as the "oxide definition" (OD) regions because over these regions only a very thin gate-oxide will be made. The hardened photoresist is left on top of the Si_3N_4 to protect it over the OD regions. Next, the Si_3N_4, wherever it is not protected by the photoresist, is removed by etching it away with a hot phosphoric acid. The SiO_2 is then removed with a hydrofluoric acid etch. Finally, the remaining photoresist is chemically removed with a process that leaves the remaining Si_3N_4 intact. The remaining Si_3N_4 will act as a mask to protect the active regions while the isolation structures are being made.

These steps result in a thin layer of thermal SiO_2, as well as a layer of silicon nitride (Si_3N_4), covering all active (OD) regions as shown in Fig. 2.5.

2.1.5 Transistor Isolation

Parasitic transistors are formed everywhere on a silicon die wherever a conductor appears above and between the junction regions of different transistors. If the electrical potential on a conductor should approach the threshold voltage of a parasitic transistor underneath it, undesired leakage currents may flow between transistors that are intended to be unconnected. In order to prevent this, extra processing is performed to ensure that these parasitic transistors can not conduct appreciable current. Two popular methods to isolate transistors are local oxidation of the silicon (LOCOS) and shallow-trench isolation (STI).

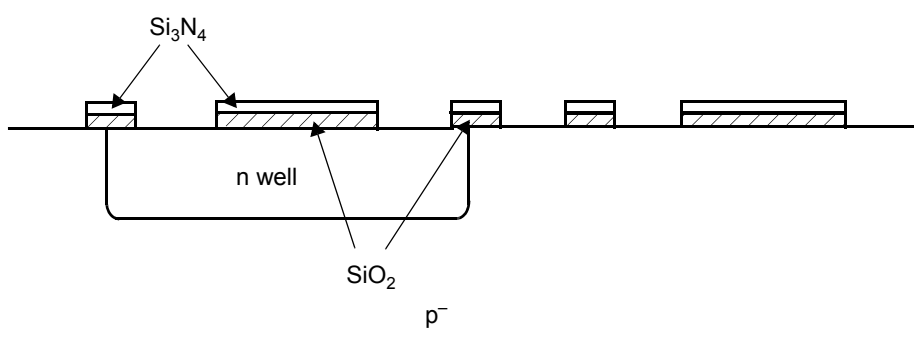


Fig. 2.5 The cross section of the wafer after the oxide definition (OD) regions are patterned.

Local Oxidation of Silicon (LOCOS)

LOCOS processing involves the implantation of additional dopants (*filed-implants*) between transistors to ensure any parasitic transistors have a very large threshold voltage, followed by the creation of very thick layers of SiO_2 (*field-oxide*) above the field-implants.

First, the *field-implants* are introduced under where the *field-oxide* will be grown. For example, boron is implanted under the field-oxide everywhere except in the n well regions. This implant guarantees that the silicon under the field-oxide will never invert (or become n) when a conductor over the field-oxide has a large voltage. For the field-oxide in the well regions, where p-channel transistors will eventually reside, an n-type implant such as arsenic (As) could be used. Often, it is not necessary to include field-implants under the field-oxide of the well regions because the heavier doping of the well (compared to that of the substrate) normally guarantees that the silicon will never invert under the field-oxide in these regions.

When implanting the field-implants in the substrate regions, it is necessary to first cover the wells with a protective photoresist, PR_3, so the n-well regions do not receive the p implant. This can be done using the same mask, M_1, that was originally used for implanting the n wells, but now a positive photoresist is used. This positive photoresist remains where the mask is opaque (i.e., dark), which corresponds to the well regions.

After the exposed photoresist has been dissolved, we now have the cross section shown in Fig. 2.6. Notice that at this step, all the active regions, where eventually the transistors will reside, are protected from the field implants by the Si_3N_4 and SiO_2. Additionally, the complete well regions are also protected by PR_3. The field-implant will be a high-energy implant with a fairly high doping level. Before the field-oxide is grown, PR_3 is removed, but the silicon-nitride–silicon-dioxide sandwich is left.

The next step is to grow the field-oxide, SiO_2. There are two different ways that SiO_2 can be grown. In a wet process, water vapor is introduced over the surface at a moderately high temperature. The water vapor diffuses into the silicon and, after some intermediate steps, reacts according to the formula

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \tag{2.1}$$

In a dry process, oxygen is introduced over the wafer, normally at a slightly higher temperature than that used in the wet process, and reacts according to the formula

$$Si + O_2 \rightarrow SiO_2 \tag{2.2}$$

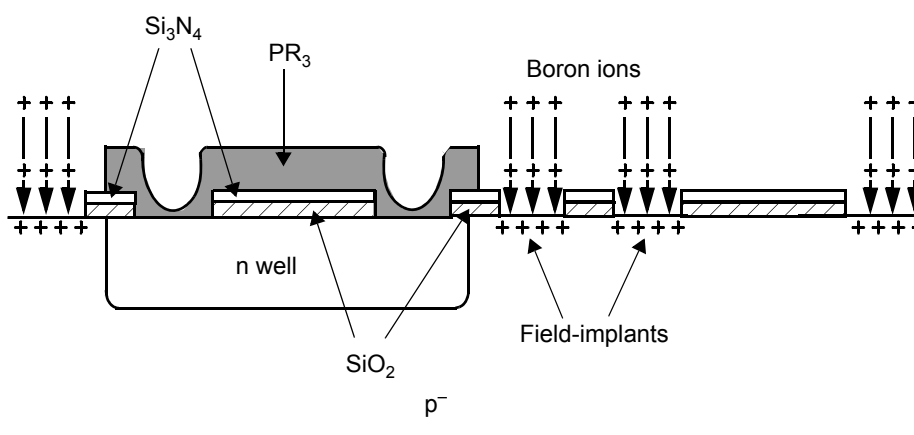


Fig. 2.6 The cross section when the field-implants are being formed in a LOCOS process.

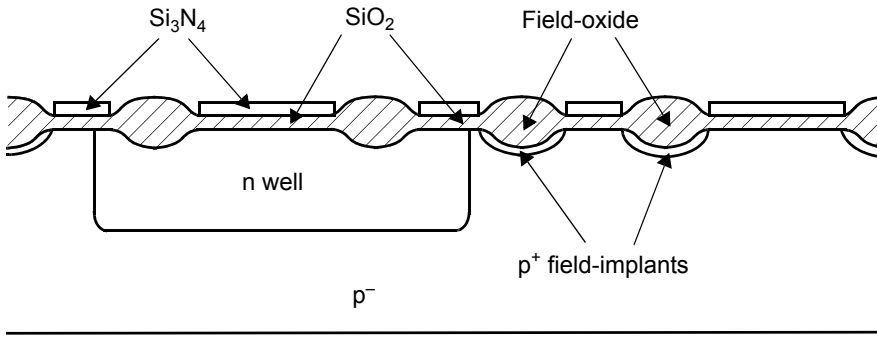


Fig. 2.7 The cross section after the field-oxide has been grown in a LOCOS process.

Since both of these processes occur at high temperatures, around 800 to 1200 °C, the oxide that results is sometimes called a *thermal oxide.*

The field oxide does not grow wherever CVD-deposited Si_3N_4 remains, because the Si_3N_4 is relatively inert to both water and oxygen. Wherever the process does occur, the volume increases because oxygen atoms have been added. Specifically, SiO_2 takes up approximately 2.2 times the volume of the original silicon. This increase will cause the SiO_2 to extend approximately 45 percent into, and 55 percent above, what previously was the surface of the silicon. The resulting cross section is shown in Fig. 2.7. Note that in our example process, the field-oxide in the substrate region has field-implants under it, whereas the field-oxide in the wells does not.

When growing thermal SiO_2, the wet process is faster because H_2O diffuses faster in silicon than O_2 does, but the dry process results in denser, higher-quality SiO_2 that is less porous. Sometimes, growing the field-oxide starts with a dry process, changes to a wet process, and finishes with a dry process. When thin layers of SiO_2 are grown, as the next section describes, usually only a dry process is used.

LOCOS is the preferred method for transistor isolation when minimum feature sizes exceed 0.25 μm. At smaller feature sizes the rounded corners of the field-oxide take up too much space and improved isolation processing is required.

Shallow-Trench Isolation (STI)

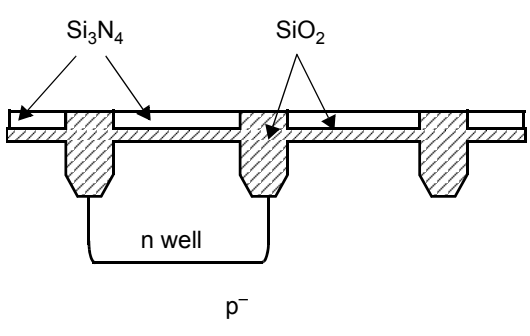


Fig. 2.8 The resulting wafer cross section when shallow-trench isolation (STI) is used between transistors.

A STI process involves etching trenches into the silicon substrate between the active regions and filling the trenches with oxide. As with the field-oxide in a LOCOS process, the trench locations are defined by a Si_3N_4 layer. Filling the trenches is a two-step process: first, the trenches are lined with a thin SiO_2 layer that is thermally grown; then additional oxide is deposited over the entire wafer, filling the trenches and leaving behind a very uneven surface. Finally, the top surface is polished to planarize it for further processing. These steps are performed at the start of the process flow, prior to well definition. An example wafer cross-section when STI is used in place of LOCOS is illustrated in Fig. 2.8. STI provides good isolation between transistors even when very closely spaced and is currently in wide use. However, it requires more

steps than LOCOS and is therefore more expensive. Moreover, the creation and filling of the trenches places a strain on the silicon wafer's lattice structure, which impacts the electrical characteristics of nearby transistors.

2.1.6 Gate-Oxide and Threshold-Voltage Adjustments

In the next step, the Si_3N_4 is removed using hot phosphoric acid. If a thin layer of SiO_2 is under the Si_3N_4, protecting the surface, as shown in Fig. 2.7, this SiO_2 is also removed, usually with hydrofluoric acid. The high-quality, thin gate-oxide is then grown using a dry process. It is grown everywhere over the wafer to a thickness of between about 1 and 30 nm.

After the gate-oxide has been grown, donors are implanted so that the final threshold voltages of the transistors are correct. Note that this implantation is performed directly through the thin gate-oxide since it now covers the entire surface. Processes differ in how they realize the threshold-adjust step.

Key Point: *Transistors are fabricated inside the "active" or "oxide definition" (OD) regions of the microcircuit. Over OD regions, only a very thin oxide separates the polysilicon gates from the transistor channel regions underneath, and additional dopants are introduced to control the threshold voltages. Surrounding the OD regions are isolation structures to prevent parasitic transistors from conducting leakage currents.*

In a simple process, the threshold voltages of both the p- and n-channel transistors are adjusted at the same time. We saw in the Appendix of Chapter 1 that the n-channel transistors require a boron implant to increase V_{tn} from its native value to its desired value. If the n wells are doped a little heavier than ideal, the native threshold voltage of the p-channel transistors in the well will also be lower than desired. As a result, the same single boron threshold-adjust implant can bring the NMOS and PMOS threshold voltages to their desired value.

By using a single threshold-voltage-adjust implant for both n-channel and p-channel transistors, two photoresist masking steps are eliminated. If the different transistors are individually implanted, then the second of two types of transistors has to be protected by, say, a negative photoresist while the first type is being implanted. Next, a positive photoresist can be used with the same mask to protect the first type of transistor while the second type is being implanted. The mask used is normally the same mask used in forming the n wells, in other words, M_1. Thus, no additional mask is required, but a number of additional processing steps are needed. The major problem with using a single threshold-adjust implant is that the doping level of the n well is higher than optimum. This higher doping level increases the junction capacitances and the body effect of the transistors in the well. Separate p- and n-type threshold adjust implants allow optimum well doping and are currently more popular. The cross section at this stage is shown in Fig. 2.9.

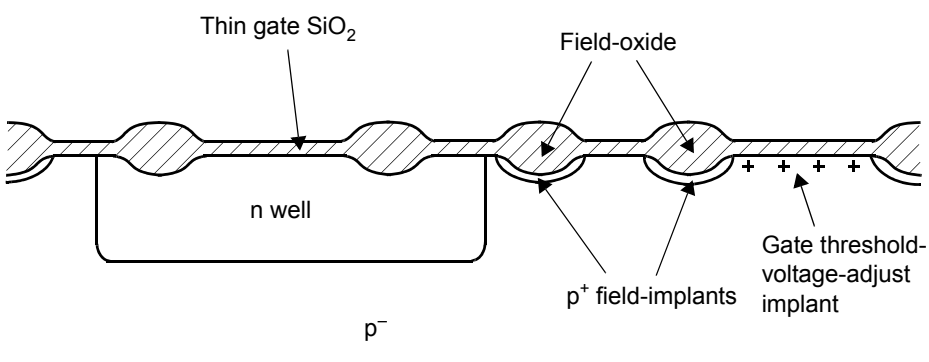


Fig. 2.9 Cross section after the thin gate-oxide growth and threshold-adjust implant.

2.1.7 Polysilicon Gate Formation

The next step in the process is chemical deposition of the polysilicon gate material. One method to create polysilicon is to heat a wafer with silane gas flowing over it so the following reaction occurs

$$SiH_4 \rightarrow Si + 2H_2 \tag{2.3}$$

If this reaction occurs at high temperatures, say, around 1000 to 1250 °C, and the original surface of the wafer was single crystal, the deposited silicon will also be single crystal. This approach is used both when epitaxial layers are grown in bipolar processes and in some modern CMOS processes. However, when depositing the polysilicon gates the original surface is SiO_2 and the wafer is heated only to about 650 °C. As a result, the silicon that is deposited is noncrystalline, or amorphous. Thus, this silicon is often referred to as *polysilicon.*

It is desirable that the polysilicon gates have low resistivity. Any series resistance in the gate will reduce the speed of the resulting transistors, and be an additional source of noise in the circuits. Hence, after the polysilicon is deposited, it is ion implanted with arsenic to increase its conductivity. A typical final resistivity for polysilicon might be 10 to 30 Ω/□, and its thickness might be around 0.25 μm. An additional step may be used to create a layer of low-resistivity salicide on top of the polysilicon gate, further reducing its resistance. For analog circuits, polysilicon strips may also be used as integrated circuit resistors rather than transistor gates. To accommodate this, an additional mask may be used to change the ion implantation and block the salicide resulting in a higher resistivity of typically 500 Ω to 3 kΩ/□ depending upon the processing details, making the material useful for realizing resistor values of 100s Ωs to 10s kΩ.

After the deposition just described, the polysilicon gate material covers the entire wafer. This polysilicon is then patterned using a new mask, M_3, and a positive photoresist, PR_4. The mask is opaque where hardened polysilicon should remain. After the nonhardened photoresist is removed, the polysilicon is etched away using a reactive plasma etch. This etch removes all of the polysilicon not protected by photoresist but removes very little of the underlying SiO_2. This thin gate-oxide layer is used to protect the surface during the next step of junction implantation. The cross section at this phase is shown in Fig. 2.10.

2.1.8 Implanting the Junctions, Depositing SiO_2, and Opening Contact Holes

The next step involves the ion implantation of the junctions. In our example process, the p^+ junctions are formed first by placing positive photoresist, PR_5, everywhere except where the p^+ regions are desired. A new mask, M_4, is

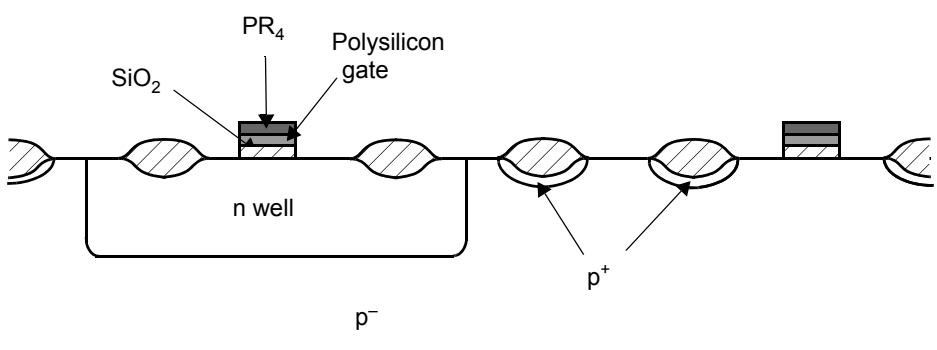


Fig. 2.10 Cross section after depositing and patterning the polysilicon gates.

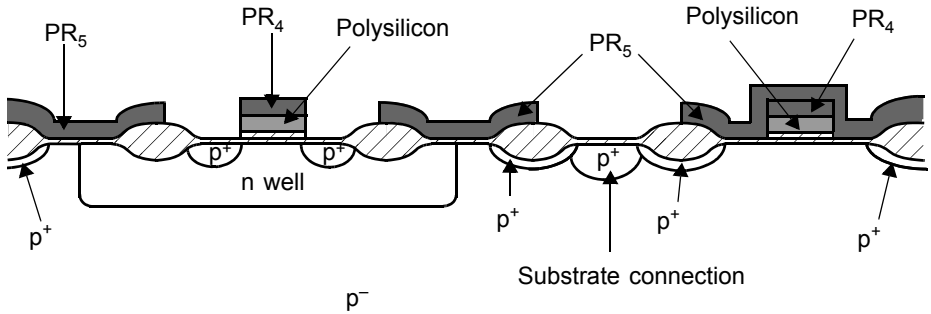


Fig. 2.11 Cross section after ion-implanting the p^+ junctions.

used in this step. The p^+ regions are then ion implanted, possibly through a thin oxide in some processes. The cross section at this stage is shown in Fig. 2.11.

Notice that the p^+ junctions of the p-channel transistors are defined on one edge by the field-oxide and, more importantly, next to the active gate area by the edge of the polysilicon gate. During the implantation of the boron, it was the gate polysilicon and the photoresist over it that protected the channel region from the p^+ implant. Thus, the p^+ junctions are *self-aligned* to the polysilicon gates, resulting in very little overlap (i.e., a small L_{ov}, as defined in Chapter 1). Also, note that the effective channel areas of the transistors are defined by the intersection of the gate-defining mask, M_3, and the mask used in defining the active regions, M_2 (i.e., the mask used in defining where Si_3N_4 remains). Thus, these are the two most important masks in any MOS process. The development of this self-aligned process has proven to be an important milestone in realizing small high-speed transistors.

> ***Key Point:*** *The edge of transistor source and drain junctions are defined by the edge of the polysilicon gate above. This "self-alignment" of the gate and junctions was key in the development of small high-speed transistors.*

Also notice that a p^+ junction has been implanted in the substrate region. This junction, called a *substrate tie,* is used to connect the substrate to ground in microcircuits. These substrate ties are liberally placed throughout the microcircuit to help prevent latch-up, a problem discussed at the end of this chapter. In addition, the underside of the wafer would normally be connected to ground as well, through a package connection.

Next, the photoresists are all removed using acetone. The p^+ active regions are then protected using the same mask, M_4, that was used for the previous step, but now a negative photoresist, PR_6, is used. The n^+ junctions are then implanted using arsenic. The cross section at the end of this stage is shown in Fig. 2.12.

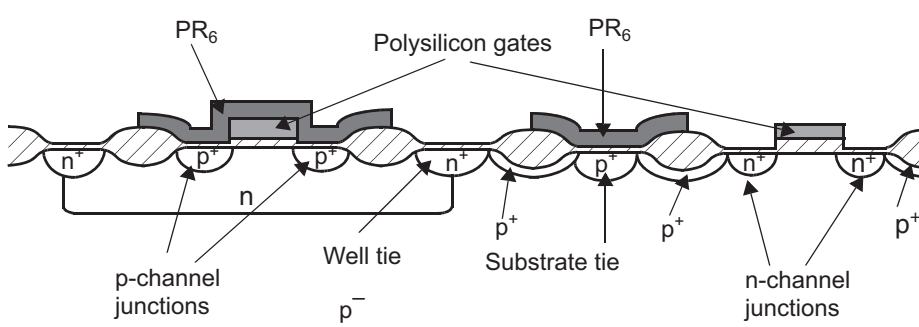


Fig. 2.12 Cross section after ion-implanting the n^+ junctions.

After the junctions have been implanted and PR_6 has been removed, the complete wafer is covered in CVD SiO_2. This protective *glass* layer can be deposited at moderately low temperatures of 500 °C or lower. The deposited SiO_2 might be 0.25 to 0.5 μm thick.

The next step is to open contact holes through the deposited SiO_2. The contact holes are defined using mask M_5 and positive resist PR_7.

2.1.9 Annealing, Depositing and Patterning Metal, and Overglass Deposition

After the first layer of CVD SiO_2 has been deposited, the wafer is annealed. As mentioned earlier in this section, annealing entails heating the wafer in an inert gas (such as nitrogen) for some period of time (say, 15 to 30 minutes) at temperatures up to 1000 °C. The resulting thermal vibrations heal the lattice damage sustained during all the ion implantations, broaden the concentration profiles of the implanted dopants, and increase the density of the deposited SiO_2.

Next, interconnect metal is deposited everywhere. Historically, aluminum (Al) has been used for the interconnect. However, other metals have been used that have less of a tendency to diffuse into the silicon during electrical operation of the microcircuit. Copper is increasingly being used to take advantage of its lower resistivity, an important consideration for very thin wires and for wires conducting large currents. The metal is deposited using evaporation techniques in a vacuum. The heat required for evaporation is normally produced by using electron-beam bombarding, or possibly ion bombarding in a sputtering system. After the metal is deposited on the entire wafer, it is patterned using mask M_6 and positive photoresist PR_8, and then it is etched.

At this time, a low-temperature annealing might take place to give better bonds between the metal and the silicon. The temperature of this annealing must be less than 550 °C so the aluminum doesn't melt.

Key Point: *Up to 10 or even more layers of metal are patterned above the silicon surface, separated by insulating oxide, to provide interconnect between all the devices in a circuit.*

Next, an additional layer of CVD SiO_2 is deposited, additional contact holes are formed using mask M_7 and photoresist PR_9, and then a second layer of metal is deposited and etched using mask M_8 and photoresist PR_{10}. Often the primary use of top layers of metal might be to distribute the power supply voltages. Lower layers would be used more often for local interconnects in gates. In modern fabrication, this process may be repeated ten or more times to provide a much denser interconnect.

After the last level of metal is deposited, a final *passivation*, or *overglass,* is deposited for protection. This layer would be CVD SiO_2, although often an additional layer of Si_3N_4 might be deposited because it is more impervious to moisture.

The final microcircuit processing step is to etch openings in the passivation to metal pads located in the top metal layer to permit electrical contacts to be formed to the circuit. This final step would use mask M_9 and photoresist PR_{11}. A cross section of the final microcircuit for our example process is shown in Fig. 2.13.

2.1.10 Additional Processing Steps

This chapter has focused primarily on an example process representative of a technology with minimum feature sizes of approximately 0.5 μm. However, many variations, often involving additional masks, are possible. Additional steps are certainly required to realize smaller feature sizes. Some of the possible variations are as follows:

1. Two wells may exist: one for p-channel transistors and one for n-channel transistors. This *twin-well process* allows both wells to be optimally doped.
2. An additional polysilicon layer may be deposited over the first layer and separated by a thin thermal oxide layer. This extra poly layer can be used to realize highly linear poly-to-poly capacitors.
3. An additional polysilicon layer might be formed that has an extremely high resistivity (say, 1 GΩ/□). This high resistivity is used to realize resistor loads in four-transistor, static random-access memory (SRAM) cells.

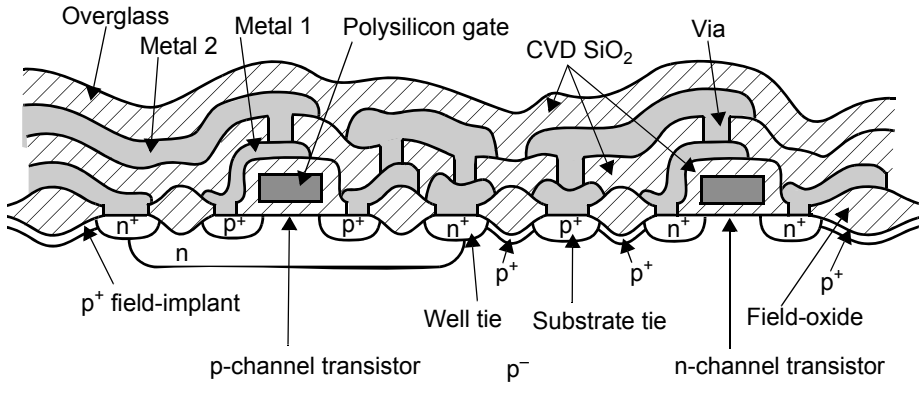


Fig. 2.13 Final cross section of an example CMOS microcircuit.

4. Field-implants may exist under the field-oxide in the well regions as well as under the field-oxide in the substrate regions.
5. Often, the n-channel and the p-channel transistors will have separate threshold-voltage-adjust implants.
6. The microcircuit might have up to ten or more layers of metal.
7. It is usually necessary to add several additional steps so that the surface is made smoother, or *planarized,* after each metal-patterning step. This is normally done by a reactive etching process in which the metal is covered with SiO_2 and the hills are etched faster than the valleys.
8. Different metals might be used for the silicon contacts than for the interconnect, to obtain better fill in and less diffusion into the silicon surface.
9. Thin-film nichrome resistors may exist under the top layer of metal.
10. Additional ion implantation and salicide blocking steps may be used to produce polysilicon strips with high sheet resistance for use as resistors.
11. The transistors may be realized in an epitaxial layer. For example, the substrate may be p^{++}, and a p^--epitaxial layer would be grown on top. The advantages of this type of wafer are that it is more immune to a destructive phenomenon called latch-up (described at the end of this chapter), and it is also more immune to gamma radiation in space. Finally, it greatly minimizes substrate noise in microcircuits that have both analog and digital circuits, (i.e., mixed-mode microcircuits).
12. "Shallow-trench isolation" involves etching trenches into the silicon substrate between transistors to reduce the coupling between them, which is increasingly problematic at small feature sizes.
13. In order to mitigate hot carrier effect, the narrow portion of the source/drain regions immediately adjacent to the channel may be ion implanted to reduce their doping levels. Unfortunately, these "lightly-doped drain" regions also introduce a significant resistance in series with the channel.
14. Impurities may be introduced into the crystalline silicon lattice in order to stretch or compress the MOSFET channel regions. For example, silicon germanium has a larger lattice spacing than pure silicon, so by using some silicon germanium to stretch the regular silicon crystal lattice, an increase in electron mobility is effected.
15. Ion implantation with very high acceleration may be used to implant dopants at a depth greater than any of the transistor features; for example, approximately 2 μm below the substrate surface. This is sometimes used to create a buried deep n well region below the silicon substrate and contacted by a regular n well. A critical circuit may be isolated from noise in neighboring circuits by enclosing it in such a buried n well making it useful for mixed analog-digital integrated circuits.

16. Additional processing steps may be used to ensure that bipolar transistors can be included in the same microcircuit as MOS transistors. This type of process is called a BiCMOS process and is particularly popular for analog high-speed microcircuits.

2.2 CMOS LAYOUT AND DESIGN RULES

It is the designer's responsibility to determine the geometry of the various masks required during processing. The process of defining the geometry of these masks is known as *layout* and is done using a CAD program. Here, we describe some typical layout design rules and the reasons for these rules.

2.2.1 Spacing Rules

When designing the layout, typically the designer does not need to produce the geometry for all of the masks because some of the masks are automatically produced by the layout program. For example, the p^+ and n^+ masks used for the source and drain regions are usually generated automatically. Also, the program might allow the designer to work in the final desired dimensions. The layout program then automatically sizes the masks to account for any lateral diffusion or etching loss; this sizing produces larger- or smaller-dimension masks. For example, a designer might draw a polysilicon line so that a transistor would have a 0.1-μm length. The program might then produce a mask that had a 0.12-μm line width. This increased mask sizing would account for the junction overlap due to lateral diffusion and the polysilicon loss due to etching.

In a modern layout program, the layout of some circuit *cells* might already be performed and stored in a library. During overall layout, these cells are then parametrically adapted to a required size, and the corresponding geometries for every layer are automatically generated. Often, when the cells are being connected, they might be automatically *placed an d routed*, or connected, by the program. The designer might then interactively modify this automatically generated layout. Thus, as time goes on, the layout becomes more automated as more cells become available. However, the designer must still take direct control of the layout of critical cells, especially when the layout must be small or the resulting circuits must be fast. For example, one would rarely allow a computer to automatically generate the layout of a memory cell where space and capacitive loading of the connecting buses are critical. Thus, a digital microcircuit designer must be knowledgeable about the design rules that govern the layout required for the process used.

The two most important masks are those for the active (OD) region and for the gate polysilicon. The intersection of these two masks becomes the channel region of MOS transistors. For example, consider Fig. 2.14(*a*), which shows a simplified view of a MOS transistor, and Fig. 2.14(*b*), which shows the corresponding layout of the active mask and the polysilicon, or poly, mask. In Fig. 2.14(*b*), the poly mask runs vertically. The length of the poly that intersects the active-region mask is the transistor width, W, and the width of the poly line is the transistor length, L, as Fig. 2.14 shows.

The design rules for laying out transistors are often expressed in terms of a quantity, λ, where λ is 1/2 the minimum permitted gate length. This generalization allows many of the design rules to be simply expressed, independent of the true value for the minimum channel length (i.e., 2λ). Figure 2.14(*b*) shows the smallest possible transistor that can be realized in a given process when a contact must be made to each junction. Also shown are many of the minimum dimensions in terms of λ.

When we express design rules in terms of λ, we assume that each mask has a worst-case alignment of under 0.75λ. Thus, we can guarantee that the relative misalignment between any two masks is under 1.5λ. If an overlap between any two regions of a microcircuit would cause a destructive short circuit, then a separation between the corresponding regions in a layout of 2λ guarantees this will never happen. For example, consider the poly mask and the contact mask in Fig. 2.14(*b*). If these two regions overlap in the microcircuit, then the

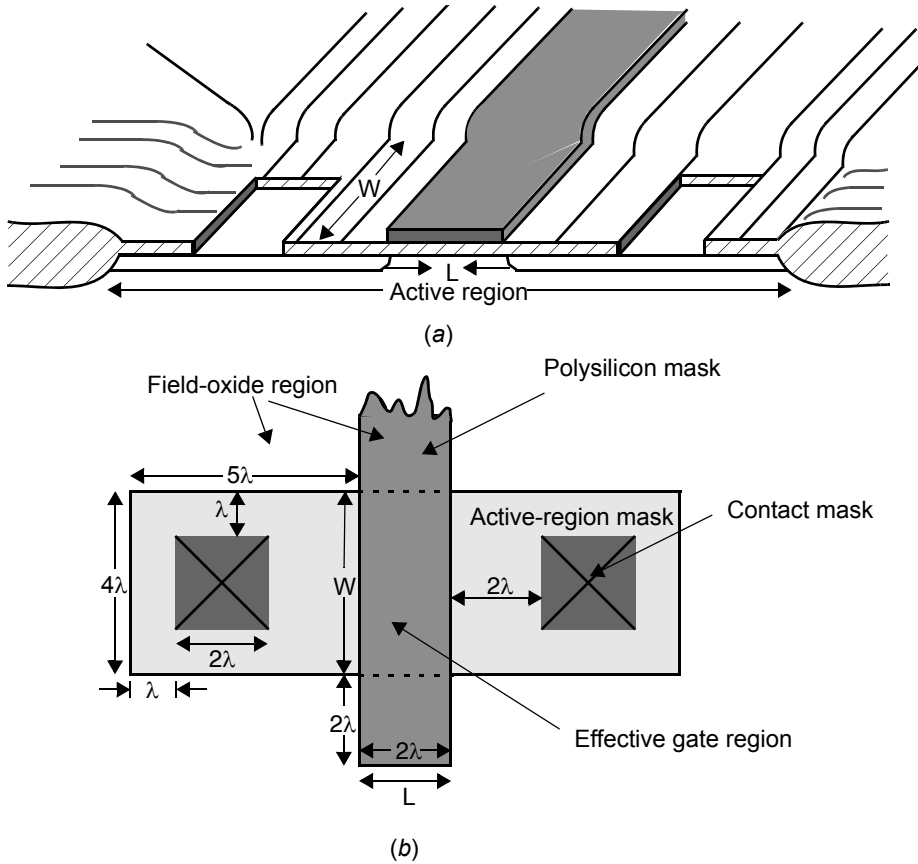


Fig. 2.14 (*a*) A simplified view of a partially finished transistor and (*b*) the corresponding layout of the active, polysilicon, and contact masks.

metal used to contact the source junction is also short-circuited to the gate poly, causing the transistor to be always turned off, as shown in Fig. 2.15. If the source happens to be connected to ground, this error also short-circuits the gate-to-ground. To prevent this type of short, the contact openings must be kept at least 2λ away from the polysilicon gates.

Another example of a catastrophic failure due to misalignment is a gate that does not fully cross the active region (also shown in Fig. 2.15). Since the junctions are implanted everywhere in the active region except under the gate, this misalignment causes a short circuit between the source and the drain—thus the design rule that polysilicon must always extend at least 2λ past the active region.

Another design rule is that active regions should surround contacts by at least 1λ. If, in reality, an overlap exists between the edge of the active-region mask and the contact mask, no disastrous shorts occur. The circuit still works correctly as long as sufficient overlap exists between the contact and the active masks so that a good connection is made between the aluminum interconnect and the junction. Since the maximum relative misalignment is 1.5λ, having the source (or drain) region surround the contact by 1λ and a minimum contact width of 2λ guarantees an overlap of at least 1.5λ.

The few design rules just described are sufficient to allow one to estimate the minimum dimensions of a junction area and perimeter before a transistor has been laid out. For example, assume that in Fig. 2.14 a contact is to

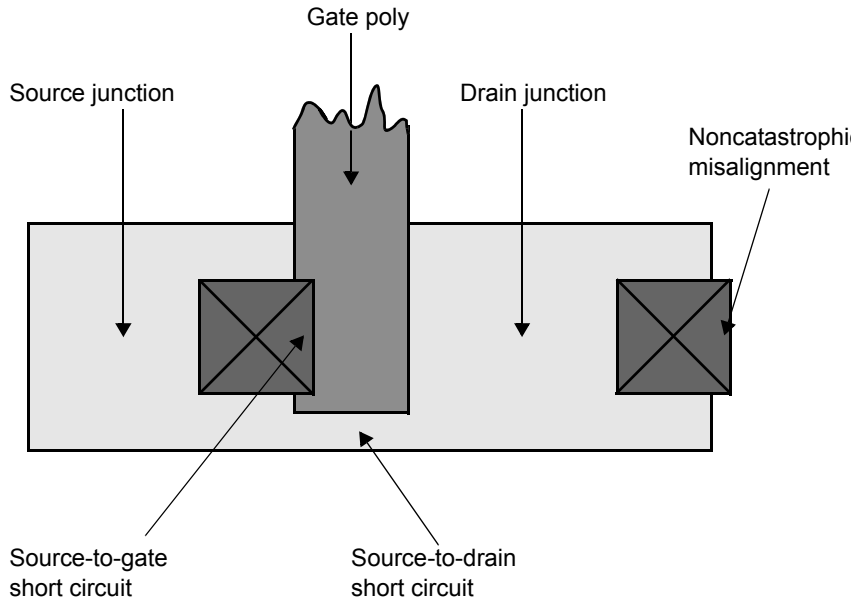


Fig. 2.15 Mask misalignment that results in catastrophic short circuits and an example of a non-catastrophic misalignment.

be made to a junction; then the active region must extend past the polysilicon region by at least 5λ. Thus, the minimum area of a small junction with a contact to it is

$$A_s = A_d = 5\lambda W \tag{2.4}$$

where W is the transistor width. Similarly, in Fig. 2.14, the perimeter of a junction[3] with a contact is given by

$$P_s = P_d = 10\lambda + W \tag{2.5}$$

These estimates may be used when estimating the parasitic capacitances in the transistor models. They may also be used in SPICE to simulate circuits so the parasitic capacitances are determined more accurately. However, note that they are only estimates; the true layout will differ somewhat from these rough estimates.

Key Point: *Finite tolerances during integrated circuit manufacturing impose constraints on the minimum sizes and spacing of transistor and interconnect features. These constraints may be expressed as multiples of λ, equal to one-half the minimum gate length. They influence parasitic capacitances, and ultimately the circuit bandwidth which an analog designer may expect.*

Sometimes, when it is important to minimize the capacitance of a junction, a single junction can be shared between two transistors. For example, consider the series connection of two transistors shown in Fig. 2.16(*a*). The active, poly, and contact masks might be laid out as shown in Fig. 2.16(*b*). Notice that a single junction is shared between transistors Q_1 and Q_2. The area, and especially the perimeter of this junction, are much smaller than those given by equations (2.4) and (2.5). Also, in a SPICE simulation, the area and perimeter should be divided by 2 when they are specified in each transistor description, since the junction is shared. Alternatively, all of the area and perimeter could be specified in one transistor description, and the area and perimeter of the other junction could be specified as zero.

Since the junction sidewall capacitance is directly proportional to the junction perimeter, and since this capacitance can be a major part of the total junction capacitance (because of the

3. Note that the perimeter does not include the edge between the junction and the active channel separating the junction and the gate because there is no field-implant along this edge, and the sidewall capacitance is therefore smaller along that edge.

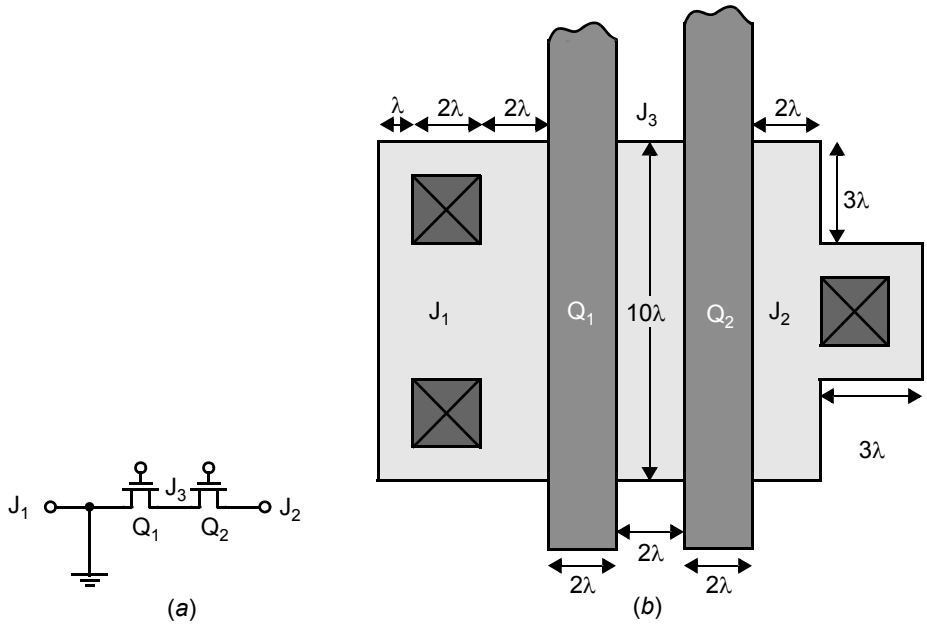


Fig. 2.16 (*a*) A series connection of two transistors and (*b*) a possible layout.

heavily doped field-implants), minimizing the perimeter is important. Note that it is impossible to share junctions between n-channel and p-channel devices as they must be located in separate substrate regions doped p- and n-type respectively.

EXAMPLE 2.1

Assuming $\lambda = 0.2\ \mu m$, find the area and perimeters of junctions J_1, J_2, and J_3 for the circuit in Fig. 2.16.

Solution

Since the width and length are shown as 10λ and 2λ, respectively, and $\lambda = 0.2\ \mu m$, the physical sizes are $W = 2\ \mu m$ and $L = 0.4\ \mu m$.

Thus, for junction J_1, using the formulas of (2.4) and (2.5), we have

$$A_{J1} = 5\lambda W = 5(0.2)2\ \mu m^2 = 2\ \mu m^2 \tag{2.6}$$

and

$$P_{J1} = 10\lambda + W = [10(0.2) + 2]\ \mu m = 4\ \mu m \tag{2.7}$$

Since this junction is connected to ground, its parasitic capacitance is unimportant and little has been done to minimize its area. Contrast this case with junction J_2, where we have

$$A_{J2} = 2\lambda W + 12\lambda^2 = 1.28\ \mu m^2 \tag{2.8}$$

The perimeter is unchanged, resulting in $P_{J2} = 4\ \mu m$. Thus, we have decreased the junction area by using the fact that the transistor is much wider than the single contact used. However, sometimes wide transistors require

additional contacts to minimize the contact impedance. For example, the two contacts used for junction J_1 result in roughly half the contact impedance of junction J_2.

Next, consider the shared junction. Here we have a junction area given by

$$A_{J3} = 2\lambda W = 0.8\ \mu m^2 \tag{2.9}$$

Since this is a shared junction, in a SPICE simulation we would use

$$A_s = A_d = \lambda W = 0.4\ \mu m^2 \tag{2.10}$$

for each of the two transistors, which is much less than 2 μm^2. The reduction in the perimeter is even more substantial. Here we have

$$P_{J3} = 4\lambda = 0.8\ \mu m \tag{2.11}$$

for the shared junction; so sharing this perimeter value over the two transistors would result in

$$P_S = P_d = 2\lambda = 0.4\ \mu m \tag{2.12}$$

for the appropriate junction of each transistor when simulating it in SPICE. This result is much less than the 4-μm perimeter for node J_1.

Because minimizing the junction capacitance is so important, one of the first steps an experienced designer takes before laying out important high-speed cells is first to identify the most critical nodes and then to investigate possible layouts that minimize the junction capacitance of these nodes.

An additional design rule has been implicitly introduced in the previous example. Notice that for junction J_2 in Fig. 2.16, part of the active region boundary is only 2λ away from the gate. This minimum junction area is the typical design rule for this case.

Several design rules are required in addition to those just mentioned. Some of these are described next, with reference to the layout of a digital inverter, shown in Fig. 2.17. Notice that the n well surrounds the p-channel active region, and therefore the p^+ junctions of the p-channel transistors, by at least 3λ. Notice also that the minimum spacing between the n well and the junctions of n-channel transistors, in the substrate, is 5λ. This large spacing is required because of the large lateral diffusion of the n well and the fact that if the n-channel junction became short-circuited to the n well, which is connected to V_{DD}, the circuit would not work. Conversely, a p^+-substrate tie can be much closer to a well because it is always connected to ground and is separated from the well by a reverse-biased junction. A typical dimension here might be 2λ. Since a p-channel junction must be inside the well by at least 3λ and an n-channel junction must be outside the well by 5λ, the closest an n-channel transistor can be placed to a p-channel transistor is 8λ.

Notice in Fig. 2.17 that metal is used to connect the junctions of the p-channel and n-channel transistors. Normally, the metal must overlap any underlying contacts by at least λ. A typical minimum width for first-level metal might be 2λ, the same as the minimum width for polysilicon. However, it can be wider as in Fig. 2.17, where it is 4λ wide.

Notice also in Fig. 2.17 that a single contact opening, known as a *butting contact*, is used to contact both the p-channel transistor source and an n^+-well tie, because both will be connected to V_{DD}. Although the outlines of the p^+ and n^+ masks are not shown in Fig. 2.17, under the contact, one half will be doped p^+ (the p-channel junction) and one half will be doped n^+ (the well tie). Also, for the n-channel transistor, a butting contact was used to connect the n-channel source to a p^+-substrate tie, and both will be connected to ground. In a typical set of design rules, a maximum distance between transistors and well (or substrate) ties is specified, and a maximum distance between substrate ties is also specified. For example, the rules might specify that no transistor can be more than 100λ from a substrate tie. These rules are necessary to prevent latch-up, a phenomenon described at the end of this chapter.

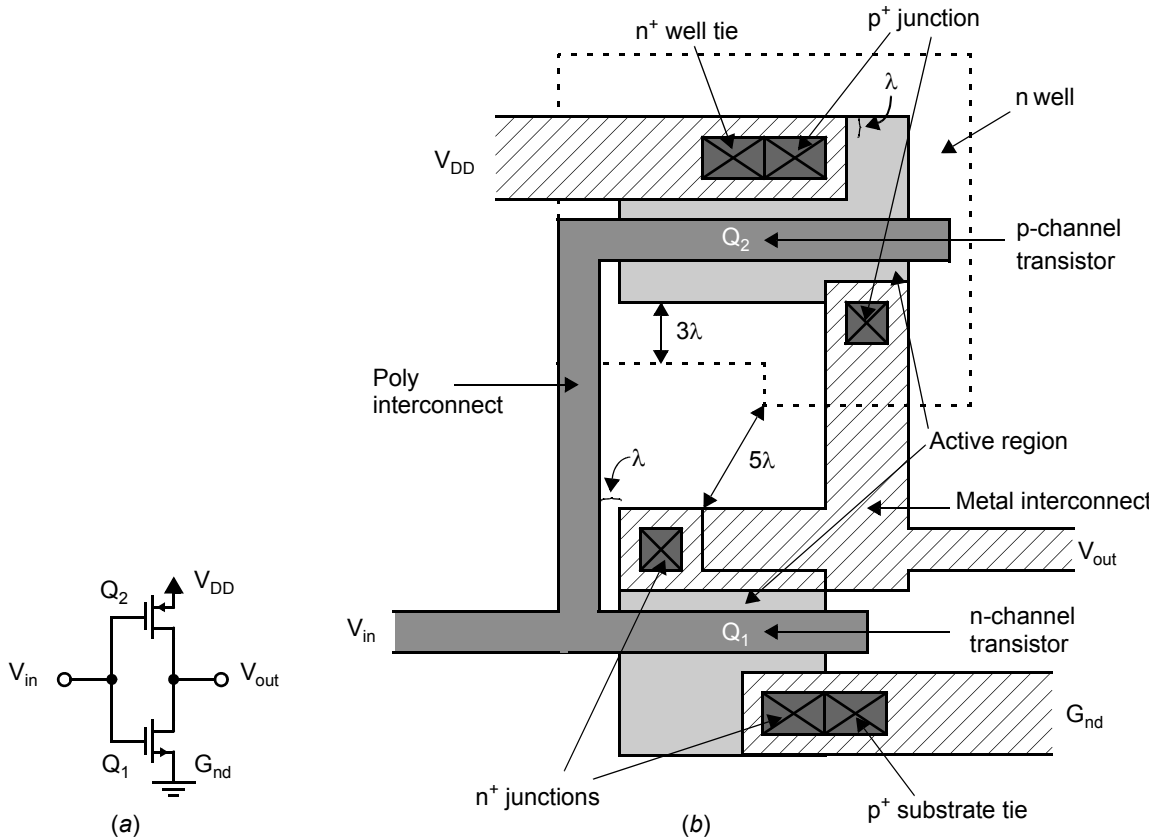


Fig. 2.17 (*a*) A CMOS digital inverter and (*b*) a possible layout with several design rules illustrated.

As a final example, we describe the layout of a large transistor. Normally, a wide transistor is composed of smaller transistors connected in parallel. This approach results in shorter individual polysilicon gate strips and, hence, a lower series gate resistance. It also reduces the junction capacitances, as we shall see in Example 2.2. A simplified layout of this approach is shown in Fig. 2.18(*a*), where four transistors that have a common gate are connected in parallel. Figure 2.18(*b*) shows the circuit corresponding to the layout in Fig. 2.18(*a*), where the transistors have been drawn in the same relative positions. Figure 2.18(*c*) shows the same circuit redrawn differently, where it is clear that the circuit consists of four transistors connected in parallel. Notice that the second and fourth junction regions are connected by metal to node 1, whereas the first, third, and fifth junction regions are connected by metal to realize node 2. Because it has a larger total junction area and especially a larger perimeter, node 2 will have a much greater junction capacitance than node 1. Thus, when the equivalent transistor is connected to a circuit, node 1 should be connected to the more critical node. Also notice the large number of contacts used to minimize the contact impedance. The use of many contacts in wide junction regions greatly minimizes voltage drops that would otherwise occur due to the relatively high resistivity of silicon junctions compared to the resistivity of the metal that overlays the junctions and connects them.[4]

4. The use of metal to overlay a higher-resistivity interconnect, such as polysilicon or heavily-doped silicon, is recommended to lower the resistivity of the interconnect.

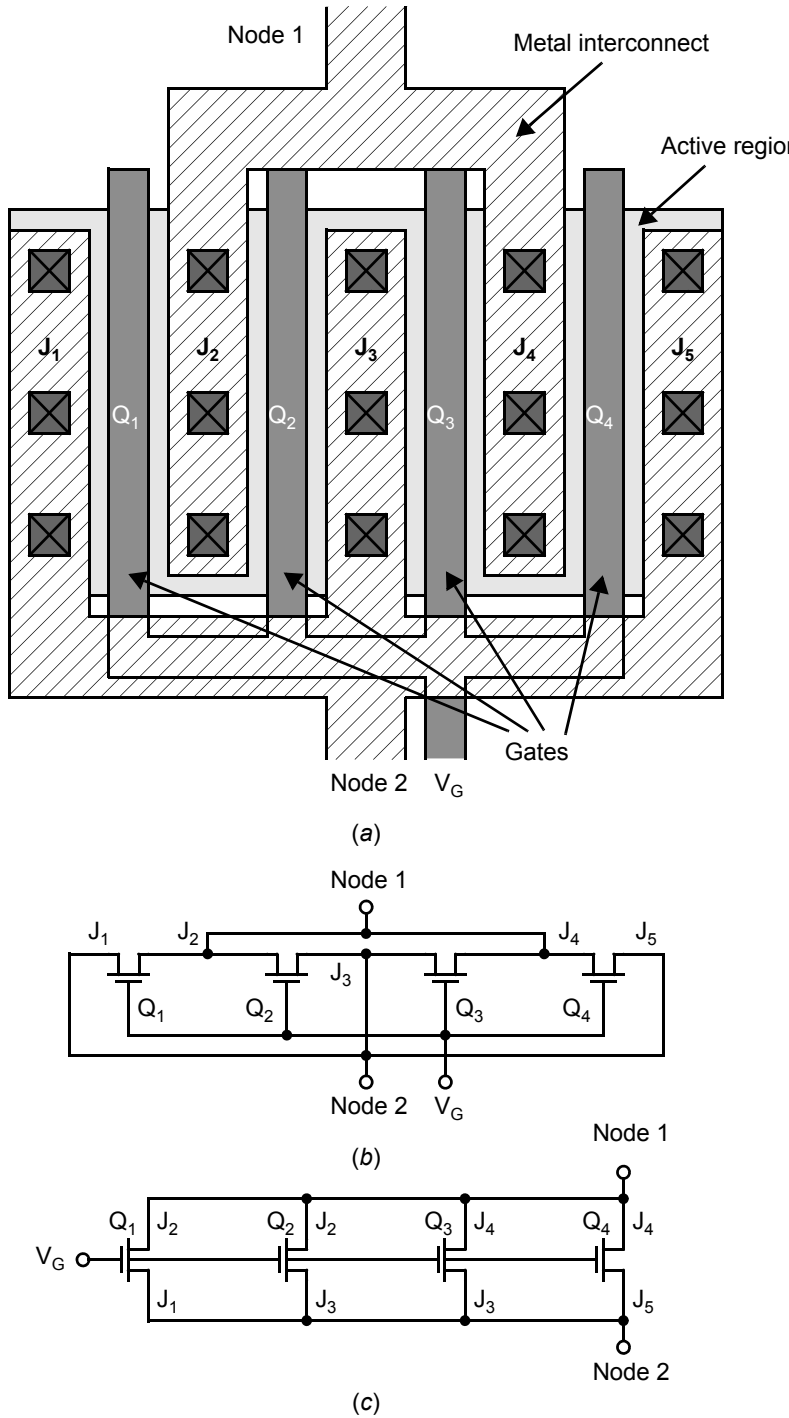


Fig. 2.18 Connecting four transistors in parallel to realize a single large transistor: (*a*) the layout, (*b*) the schematic drawn in the same relative positions as the layout, and (*c*) the circuit redrawn to make the parallel transistors more obvious.

Design rules also specify the minimum pitch between polysilicon interconnects, metal 1 interconnects, and metal 2 interconnects. These might be 2λ, 2λ, and 3λ, respectively. Metal 2 requires a larger minimum pitch because it resides further from the silicon surface where the topography is less even. The minimum widths of poly, metal 1, and metal 2 might also be 2λ, 2λ, and 3λ, respectively.

This concludes our brief introduction to layout and design rules. In a modern process, many more design rules are used than those just described. However, the reasons for using and the methods of applying these rules is similar to that which has been described. Finally, note that when one does modern integrated circuit layout, the design rules are usually available to the layout CAD program and are automatically checked as layout progresses.

EXAMPLE 2.2

Consider the transistor shown in Fig. 2.18, where the total width of the four parallel transistors is 80λ, its length is 2λ, and λ = 0.2 μm. Assuming node 2 is the source, node 1 is the drain, and the device is in the active region, find the source-bulk and drain-bulk capacitances given the parameters C_j = 0.24 fF/μm² and $C_{j\text{-}sw}$ = 0.2 fF/μm. Also find the equivalent capacitances if the transistor were realized as a single device with source and drain contacts still evenly placed.

Solution

Starting with node 1, the drain, we find that the areas of the junctions are equal to

$$A_{J2} = A_{J4} = 6\lambda \times 20\lambda = 120\lambda^2 = 4.8\ \mu m^2$$

Ignoring the gate side, the perimeters are given by

$$P_{J2} = P_{J4} = 6\lambda + 6\lambda = 12\lambda = 2.4\ \mu m$$

As a result, C_{db} can be estimated to be

$$C_{db} = 2(A_{J2}C_j + P_{J2}C_{j\text{-}sw}) = 3.3\ fF$$

For node 2, the source, we have

$$A_{J1} = A_{J5} = 5\lambda \times 20\lambda = 100\lambda^2 = 4\ \mu m^2$$

and

$$A_{J3} = A_{J2} = 4.8\ \mu m^2$$

The perimeters are found to be

$$P_{J1} = P_{J5} = 5\lambda + 5\lambda + 20\lambda = 30\lambda = 6\ \mu m$$

and

$$P_{J3} = P_{J2} = 2.4\ \mu m$$

resulting in an estimate for C_{sb} of

$$\begin{aligned} C_{sb} &= (A_{J1} + A_{J3} + A_{J5} + WL)C_j + (P_{J1} + P_{J3} + P_{J5})C_{j\text{-}sw} \\ &= (19.2\ \mu m^2)0.24\ fF/\mu m^2 + (14.4\ \mu m)0.2\ fF/\mu m \\ &= 7.5\ fF \end{aligned}$$

It should be noted that, even without the additional capacitance due to the WL gate area, node 1 has less capacitance than node 2 since it has less area and perimeter.

In the case where the transistor is a single wide device, rather than four transistors in parallel, we find

$$A_J = 5\lambda \times 80\lambda = 400\lambda^2 = 16\ \mu m^2$$

and

$$P_J = 5\lambda + 5\lambda + 80\lambda = 90\lambda = 18\ \mu m$$

resulting in $C_{db} = 7.4$ fF and $C_{sb} = 9.0$ fF. Note that in this case, C_{db} is nearly twice what it is when four parallel transistors are used.

2.2.2 Planarity and Fill Requirements

Many aspects of the fabrication process require the circuit surface to be very planar. Generally, the optics used to achieve fine lithographic resolution in modern CMOS circuits, also ensure a very narrow depth of field for the lithography. Hence, any surface roughness will blur the resulting patterns. As illustrated in Fig. 2.13, depending upon the density of metal, contacts, and polysilicon, the thickness of a microcircuit can vary considerably. Hence, it is typically required that the fraction of an overall microcircuit covered by particular layers be constrained within some range. For example, it might be required that on the first layer of metal, between 10% and 35% of the entire area of a layout be filled. Since more advanced CMOS processes require more precise planarity of the circuit, these fill requirements become more stringent. Modern processes place requirements not only on the overall circuit, but also on any particular region (or "tile") of a circuit.

In analog circuits, following the minimum fill design rules can be difficult. Analog layouts are often dominated by relatively large passive components such as resistors and capacitors, which leave many metal layers unused. A typical solution is to add superfluous "dummy" metal, polysilicon, etc. to the layout. This process is automated by many CAD tools, but analog designers may wish to tightly control or at least monitor the process to ensure the resulting dummy fill does not introduce any undesirable parasitic effects.

2.2.3 Antenna Rules

Antennal rules are intended to prevent a microcircuit from being permanently damaged during manufacture by static charges that develop on conductors in the circuit. An example is illustrated in Fig. 2.19. During the manufacturing process, before the creation of Metal 2, the circuit shown in Fig. 2.19(*a*) will look like Fig. 2.19(*b*). As it

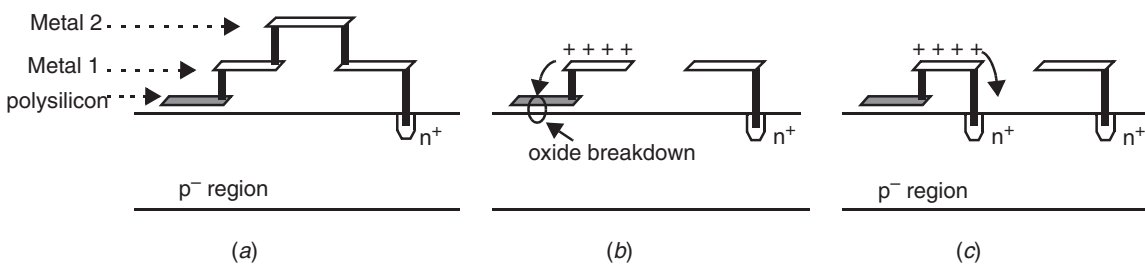


Fig. 2.19 During the manufacturing process, before the creation of Metal 2, the circuit cross-section shown in (*a*) will look like (*b*). If a large static charge develops on the polysilicon gate at this time, the gate oxide can be damaged. Antenna rules ensure that nodes at risk of sustaining this type of damage are connected to a diode to provide a discharge path, as shown in (*c*).

passes through the various fabrication steps, a significant static electric charge may develop on the node connected to the polysilicon gate. Since this node is completely isolated, the charge gives rise to static electric fields, which can have high intensity across the very thin gate oxide. If sufficient charge builds up, the oxide can break down. Antenna rules ensure that nodes at risk of sustaining this type of damage are connected to a diode from the very start of the manufacturing process. The diode is reverse biased during circuit operation, and therefore has no effect other than introducing some junction capacitance, but it provides a discharge path for any charge that may accumulate during manufacture, as shown in Fig. 2.19(*c*).

2.2.4 Latch-Up

Latch-up is a destructive phenomenon in CMOS integrated circuits that can occur when there are relatively large substrate or well currents or, equivalently, large substrate or well voltage drops, that might be caused by capacitive coupling. These triggering voltage drops often occur when power is first applied to a CMOS integrated circuit.

A latched-up circuit is equivalent to a turned-on *silicon-controlled rectifier* (SCR) between the power supply and ground. This SCR effectively short-circuits the power supplies on the microcircuit and, unless the supply-current is limited, irreparable damage will probably occur (such as a fused open bonding wire or interconnect line).

To understand latch-up, consider the cross section of the CMOS inverter shown in Fig. 2.20 with the parasitic bipolar transistors Q_1 and Q_2. Transistor Q_1 is a lateral npn, with the base being formed by the p^- substrate, whereas Q_2 is a vertical pnp, with the base being formed by the n-well region. The parasitic bipolar circuit has been redrawn in Fig. 2.21 along with some of the parasitic resistances due to the lightly doped substrate and well regions. The circuit realizes two cross-coupled common-emitter amplifiers in a positive feedback loop. This is the equivalent circuit of an SCR, which is sometimes referred to as a crowbar switch.

Normally, the parasitic bipolar transistors are off, and the voltages are as shown in Fig. 2.21(*a*). However, if latch-up is somehow triggered, they turn on when the loop gain is larger than unity, and as a result, the voltages are approximately those shown in Fig. 2.21(*b*). This turned-on SCR effectively places a short-circuit across the power-supply voltage and pulls V_{DD} down to approximately 0.9 V. If the power supply does not have a current limit, then excessive current will flow and some portion of the microcircuit may be destroyed.

To prevent latch-up, the loop gain of the cross-coupled bipolar inverters is kept less than unity primarily by having low-impedance paths from the power supplies to the substrate and well resulting in low R_n and R_p. Hence, with an n-well technology, the design rules normally specify a maximum distance between any place in the n-channel region of the microcircuit and the closest p^+ junction, which connects the substrate to ground.

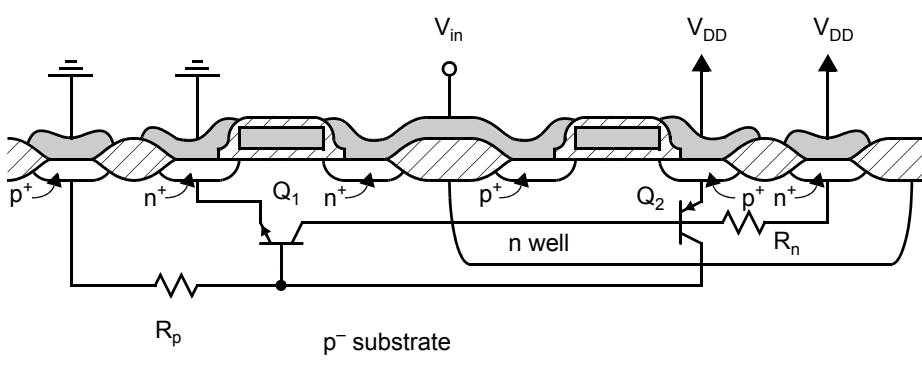


Fig. 2.20 Cross section of a CMOS inverter with superimposed schematic of the parasitic transistors responsible for the latch-up mechanism.

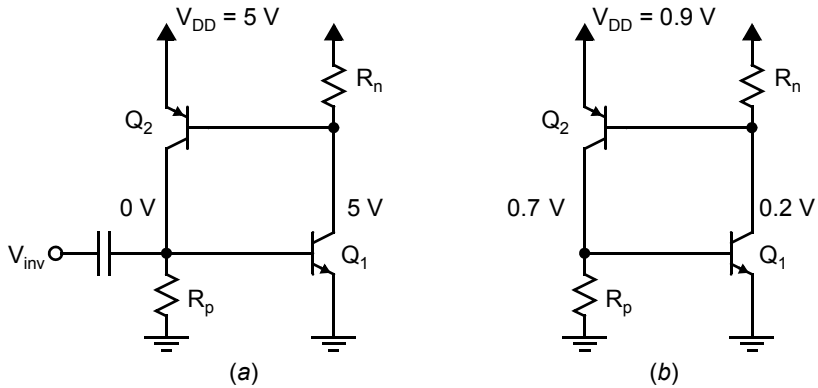


Fig. 2.21 (*a*) The equivalent circuit of the parasitic bipolar transistors, and (*b*) the voltages after latch-up has occurred.

Similarly, in the p-channel regions, the maximum distance to the nearest n^+ junction, which connects the n wells to V_{DD}, is specified. In addition, any transistors that conduct large currents are usually surrounded by guard rings, as illustrated in Fig. 2.29. These guard rings are connections to the substrate for n-channel transistors, or to the n well for p-channel transistors, that completely surround the high-current transistors. Also, ensuring that the back of the die is connected to ground through a eutectic gold bond to the package header is helpful.

One of the best ways of preventing latch-up is to use an epitaxial process, especially one with highly doped buried layers. For example, if there is a p^{++} substrate underneath the p^- epitaxial layer in which the transistors are placed, device performance is only marginally affected but the highly conductive p^{++} substrate has very little impedance to ground contacts and to the package header.

2.3 VARIABILITY AND MISMATCH

When integrated circuits are manufactured, a variety of effects cause the effective sizes and electrical properties of the components to differ from those intended by the designer. We may categorize these effects as being either systematic variations, process variations, or random variations.

2.3.1 Systematic Variations Including Proximity Effects

When lithographic techniques are used, a variety of two-dimensional effects can cause the effective sizes of the components to differ from the sizes of the glass layout masks. Some examples of these effects are illustrated in Fig. 2.22.

For example, Fig. 2.22(*a*) shows how an effective well area will typically be larger than its mask due to the lateral diffusion that occurs not just during ion implantation but also during later high-temperature steps, such as annealing. Another effect, known as *overetching,* occurs when layers such as polysilicon or metal are being etched. Figure 2.22(*b*), for example, shows overetching that occurs under the

Key Point: *Systematic variations are those observed repeatedly and consistently, even when a circuit is being mass-produced. Generally, these may be avoided by proper layout techniques, although doing so may impose some penalty in terms of layout size or performance. Process variations are observed because the conditions of manufacture (temperature, concentration levels, etc.) can never be maintained precisely constant. These manifest as device parameters that differ from one sample of a circuit to the next. Random variations are statistical in nature and are present in every individual device. Hence, they may be observed as mismatch between two identically specified transistors fabricated under the same nominal conditions.*

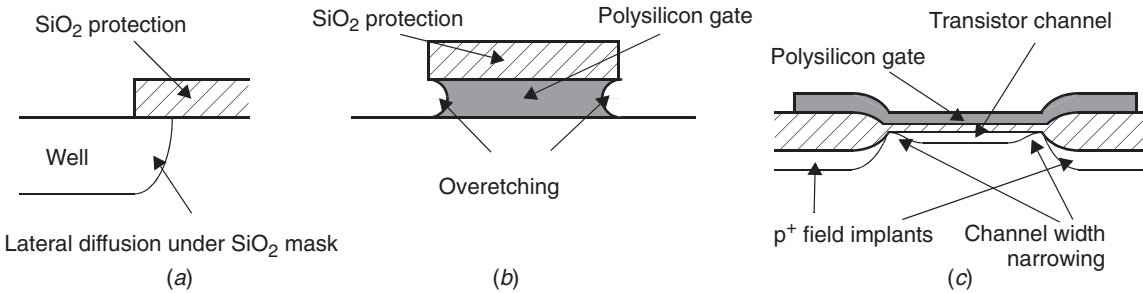


Fig. 2.22 Various two-dimensional effects causing sizes of realized microcircuit components to differ from sizes of layout masks.

SiO_2 protective layer at the polysilicon edges and causes the polysilicon layer to be smaller than the corresponding mask layout. A third effect is shown in Fig. 2.22(*c*), where an n -channel transistor is shown as we look along the channel from the drain to the source. The width of the transistor is defined by the width of the active region (as opposed to the width of the polysilicon line), and this width is determined by the separation of the isolation oxide between transistors (i.e. the field-oxide in a LOCOS process). The p^+ field implant under the field-oxide causes the effective substrate doping to be greater at the sides of the transistors than elsewhere. This increased doping raises the effective transistor threshold voltage near the sides of the transistors and therefore decreases the channel-charge density at the edges. The result is that the effective width of the transistor is less than the width drawn on the layout mask.

The features surrounding a device can also impact its electrical performance. For example, neighboring conductors give rise to parasitic capacitances in a circuit. However, there are some more subtle proximity effects. For example, when ion implantation is used for well formation, some incident atoms will scatter when striking near the edge of the photoresist, as shown in Fig. 2.23. As a result, the dopant concentration at the surface of the n-well is elevated near the well edge, gradually decreasing over a distance of 1 μm or more [Drennan, 2006]. Hence, transistors will have threshold voltages that vary considerably depending upon their location and orientation relative to

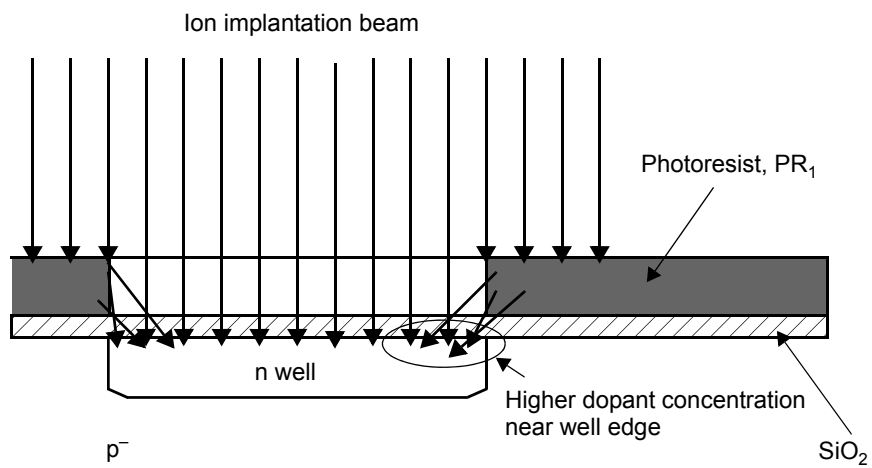


Fig. 2.23 The well edge proximity effect is caused by scattering during ion implantation.

a well edge. The conservative analog designer may layout all transistors some minimum distance from any well edge, perhaps as much as 3 μm, to avoid this effect—a huge distance in manufacturing technologies capable of feature sizes smaller than 100 nm.

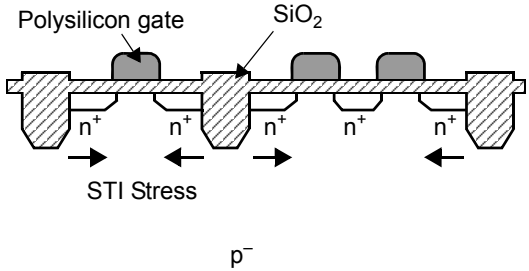


Fig. 2.24 Shallow-trench isolation (STI) places stress on the surrounding silicon effecting nearby transistor parameters.

Shallow-trench isolation can also impact the electrical properties of the surrounding silicon. The trench formation and filling places a compressive stress on the silicon lattice, as shown in Fig. 2.24. This compressive stress reduces electron mobility and increases hole mobility. It can also impact the rate of dopant diffusion, thereby influencing threshold voltage as well. Since the stress effects only those devices near the edge of the active OD regions, it can lead to mismatch between transistors that are not laid out identically. For example, if multiple transistors share the same active region they must be oriented so that they are both equidistant from the edge of the active region, and so that they both have the same orientation with respect to the edge (e.g., they should both have their source closer to the edge of the active region). A conservative approach is to always include dummy transistor structures close to the edge every active region, so that no active analog transistors are located close to an STI trench. In the most modern processes, two dummies may be required on each edge of the active regions [Drennan, 2006].

Key Point: *The absolute sizes and electrical parameters of integrated circuit components can seldom be accurately determined. For best accuracy, larger objects are made out of several unit-sized components connected together, and the boundary conditions around all objects should be matched, even when this means adding extra unused components.*

These examples illustrate typical systematic effects, but many other second-order effects influence the realized components. These other effects include those caused by boundary conditions of an object, the size of the opening in a protective layout through which etching occurs, and the unevenness of the surface of the microcircuit [Maloberti, 1994]. For these reasons, the absolute sizes and electrical parameters of integrated circuit components can seldom be accurately determined. For best accuracy, larger objects are made out of several unit-sized components connected together, and the boundary conditions around all objects should be matched, even when this means adding extra unused components. Inaccuracies also affect the ratios of sizes when the ratio is not unity, although to a lesser degree.

2.3.2 Process Variations

Variations in the manufacturing process also influence device performance. For example, the temperatures under which various fabrication steps are performed and the concentration of elements introduced will influence transistor parameters. In spite of intense effort to minimize these variations, they persist and are significant. For example, the oxide thickness may vary by 5% and dopant concentrations by 10% [Tsividis, 2002]. These translate into device parameter variations equivalent to, for example, 100 mV changes in V_{t0}, 15% for K', and 5% for junction capacitances. All designs will be subject to these process variations regardless of how much care is taken to eliminate the systematic variations described above.

To see the effect of process variations during design, several different device models are used for the analysis and simulation of any analog circuit. Each model represents a different combination of device parameter variations that may be expected to occur during mass production of the design. For example, since PMOS and NMOS devices often are subjected to different channel implantation steps, process variations may cause the PMOS and NMOS threshold voltages to vary either up or down independently of one another. Taking into account variations in all process parameters quickly leads to a huge number of permutations. In addition, a practical design will be

expected to operate effectively over a range of temperatures (which in turn influences carrier mobility and other device parameters), and over a range of different supply voltages. Collectively, these are referred to as PVT (process–voltage–temperature) variations, and are often a bane to analog designers.

EXAMPLE 2.3

Consider a NMOS transistor biased with $V_{GS} = 0.65$ V and with the device parameters listed for the 0.18-μm CMOS process in Table 1.5. How much does the drain current change with a 100 mV increase in V_{tn}, 5% decrease in C_{ox}, and 10% decrease in μ_n?

SPICE! *This may be simulated using the corner models on the text website.*

Solution

Note that since Table 1.5 indicates a threshold voltage of $V_{tn} = 0.45$ V, a V_t increase of 100 mV represents a 50% decrease in V_{eff}, from 200 mV to 100 mV. Adopting a simple square law device model, a great simplification considering the large process variations present on λ, the nominal device drain current is given by

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{eff}^2$$

Accounting for all of the process variations, the new drain current will be

$$I_{D,\,new} = \frac{1}{2}(0.9\mu_n)(0.95C_{ox})\frac{W}{L}(0.5V_{eff})^2 = (0.9 \cdot 0.95 \cdot 0.5^2)I_D = 0.214 \cdot I_D$$

representing a 79% decrease in drain current. Much of this is a result of the V_t variation.

Key Point: *Practical variations in transistor parameters, supply voltage, and temperature (PVT variations) can be a bane to analog designers trying to reliably meet specific performance criteria.*

The combination of parameter variations considered in Example 2.3 is colloquially referred to as a "slow process corner" because those variations, combined with larger-than-expected junction capacitances, result in digital circuits that are slower than what may be nominally expected from a particular manufacturing process. Designers also consider a "fast" process corner exhibiting reduced values of V_t and junction capacitances and increased C_{ox} and μ. These slow and fast corners, when combined with high and low operating temperatures, and low and high supply voltages respectively, are often considered the extreme cases under which a digital circuit must operate. Unfortunately, analog performance metrics may actually be worst under some different, but equally likely, combination of process parameter variations.

2.3.3 Random Variations and Mismatch

Even in the absence of any process variations, there are limits on the accuracy with which devices can be fabricated. For example, the number of dopants present in the channel region of modern minimum-size transistors may be as few as 100. It is impossible to ensure that two transistors will have the exact same number and location of dopants, even when they have identical layouts and are fabricated under identical conditions. As the number of dopants or their relative location varies randomly, so do transistor parameters such as threshold voltage. Channel dopants are just one source of uncertainty—there are several such factors that contribute to random variations in device parameters.

A fundamental analysis of the statistical variation of device parameters on a wafer [Pelgrom, 1989] predicts a Gaussian distribution with variance

$$\sigma^2(\Delta P) = \frac{A_P^2}{WL} + S_P^2 D^2 \tag{2.13}$$

where ΔP is the difference in some device parameter P between two devices spaced a distance D apart, W and L are the dimensions of the device, and A_p and S_p are proportionality constants usually obtained from experimental measurements. Equation (2.13) captures only random variations in device parameters, not the systematic or process variations already described. The expression is general and can be applied to transistors, capacitors, resistors, etc. although it manifests itself slightly differently in different model parameters. For example, sometimes the variation is in absolute terms, and sometimes in relative terms.

The most important random variations in a MOSFET may be modelled by variations in V_t (threshold voltage) and $K' = \mu_n C_{ox}(W/L)$.

$$\sigma^2(V_{t0}) = \frac{A_{Vt0}^2}{WL} + S_{Vt0}^2 D^2 \tag{2.14}$$

$$\frac{\sigma^2(K')}{K'^2} = \frac{A_{K'}^2}{WL} + S_{K'}^2 D^2 \tag{2.15}$$

Although there is some debate in this regard, it is generally conservatively assumed that statistical variations in V_t and K' are uncorrelated. These expressions clearly illustrate the need to closely space devices whose mismatch is of concern to us. Prime examples are differentially paired transistors and current mirroring transistors. Assuming this is done, the area dependent term will dominate and is therefore our focus.

Mismatch in Transistors with the Same V_{GS}

Transistors biased with the same gate-source voltage will have currents that vary as follows [Kinget, 2005]:

$$\left(\frac{\sigma(\Delta I_D)}{I_D}\right)^2 = \left(\frac{g_m}{I_D}\right)^2 \sigma^2(\Delta V_t) + \frac{\sigma^2(\Delta K')}{K'^2} \tag{2.16}$$

Key Point: *The parameters of devices that are closely spaced with identical layout exhibit a random variance inversely proportional to device area. In practice, the most important random variations in most analog circuits are V_t mismatch.*

This does not presume a square law and is therefore valid over all operating regions. Since both ΔV_T and $\Delta K'$ are inversely proportional to the square root of device area, the overall relative current mismatch is also inversely proportional to the square root of the device area. Substituting (2.14) and (2.15) into (2.16) and assuming closely-spaced devices so we may neglect the terms with $D \approx 0$,

$$\left(\frac{\sigma(\Delta I_D)}{I_D}\right)^2 = \frac{1}{WL}\left[A_{Vt0}^2\left(\frac{g_m}{I_D}\right)^2 + A_{K'}^2\right] \tag{2.17}$$

Hence if we want to improve the current matching by a factor of 2, we must quadruple the device area. This can be achieved, for example, by doubling both W and L.

Fig. 2.25 depicts a simple current mirror circuit where $V_{GS,1} = V_{GS,2}$. The currents will therefore vary with a standard deviation given by (2.17). Under a square law, $g_m/I_D = 1/(2V_{eff})$, so the first terms in (2.16) and (2.17) decrease with increasing V_{eff}. In subthreshold, the first term approaches a constant maximum value of $g_m/I_D = q/nkT$. A plot of current mismatch representative of closely spaced 2 μm/0.2 μm devices with the same V_{GS} in a 0.18-μm CMOS process is presented in Fig. 2.26. Clearly, V_t-mismatch is the dominant source of error except at impractically large values of $V_{eff} = V_{GS} - V_t$. This remains true even for large-area devices since both contributions are inversely proportional to device area, WL.

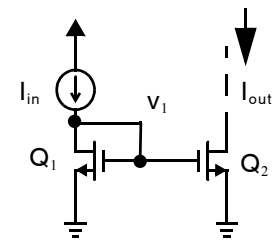


Fig. 2.25 A simple current mirror with mismatch.

EXAMPLE 2.4

Assuming a current mirror is to be biased with V_{eff}=0.4V, how large must the devices be in order to ensure the current mismatch has a standard deviation better than 1%? Assume that $A_{Vt0} = 4\ \text{mV} \cdot \mu\text{m}$ and $A_{K'} = 0.01\ \mu\text{m}$.

Solution

From Fig. 2.26, the 2 μm/0.2 μm device has a standard deviation of 3.2% at $V_{eff} = 0.4$ V. This will decrease with $\sqrt{WL}$. Hence the device must be made $3.2^2 = 10.2$ times larger in area. For example, each device could be sized 6.5 μm/0.65 μm.

Fig. 2.26 considers two transistors of fixed size with the same value of V_{GS} and a drain current I_D that varies with V_{eff}, in which case the best matching is obtained by selecting a large value of V_{eff} where the impact of V_t

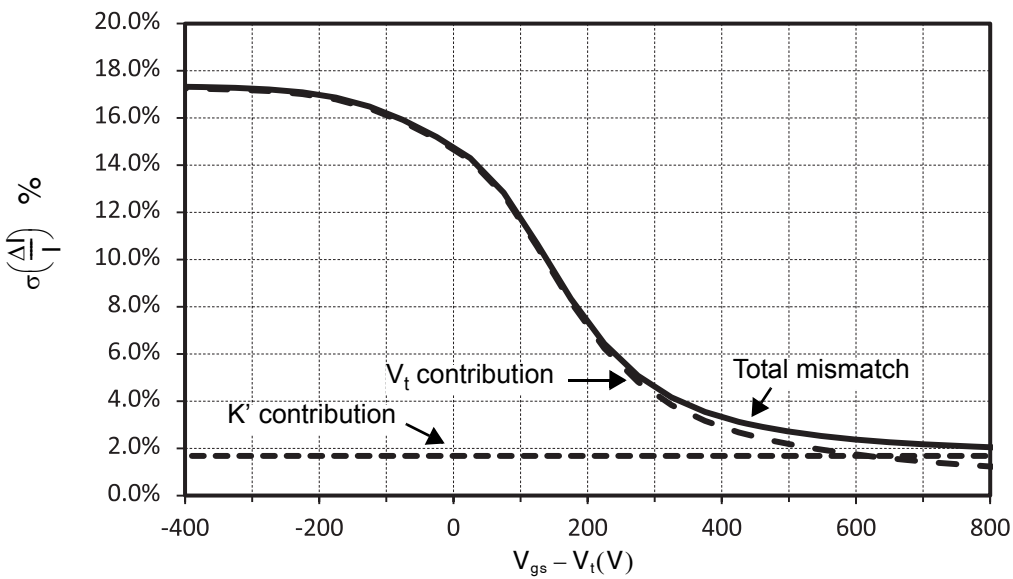


Fig. 2.26 A plot of current mismatch versus effective gate-source voltage for a simple current mirror. The values are representative of what one might see in a 0.18 μm CMOS process with $A_{Vt0} = 4\ \text{mV} \cdot \mu\text{m}$ and $A_{K'} = 0.01\ \mu\text{m}$ for devices sized $W/L = 2\ \mu\text{m}/0.2\ \mu\text{m}$.

variations is reduced. However, if instead the drain current I_D is fixed and V_{eff} varies with device aspect ratio (W/L), then the best matching is obtained at low values of V_{eff} since that implies the largest devices.

Mismatch in Differential Pairs

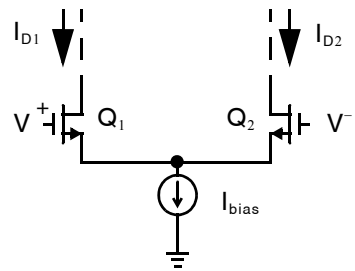


Fig. 2.27 A simple NMOS differential pair.

Consider the differential pair in Fig. 2.27 where transistors Q_1 and Q_2 are subject to random variations. When the differential pair input is zero, $V^+ = V^-$ and $V_{GS,1} = V_{GS,2}$, the currents will vary with a standard deviation given by (2.17). Hence the input offset voltage is given by

$$\sigma^2(V_{OS}) = \frac{\sigma^2(\Delta I_D)}{g_m^2} = \frac{1}{WL}\left[A_{Vt0}{}^2 + \left(\frac{I_D}{g_m}\right)^2 A_{K'}{}^2\right] \qquad (2.18)$$

The first term will practically always dominate, particularly at the input of an opamp where the differential pair devices are sized to have high transconductance per unit of current, g_m/I_D.

EXAMPLE 2.5

If a differential pair is to be biased with I_{bias}=200 μ A, how large must the devices be sized to ensure the input offset is less than 1 mV 99.8% of the time? Assume that $A_{Vt0} = 4\ \text{mV} \cdot \mu\text{m}$ and $A_{K'} = 0.01\ \mu\text{m}$.

Solution

The specs require the input offset to have a standard deviation better than $1\ \text{mV}/3 = 0.333\ \text{mV}$. Each device has a nominal drain current of I_D=100 μ A. Assuming that the first term is dominant, equation (2.18) gives

$$(0.333\ \text{mV})^2 \cong \frac{(4\ \text{mV} \cdot \mu\text{m})^2}{WL} \Rightarrow WL = 144\ \mu\text{m}^2$$

For example, if the gate length is $L = 0.5\ \mu\text{m}$, the device widths must be $W = 288\ \mu\text{m}$.

Mismatch in Transistors with Same Currents

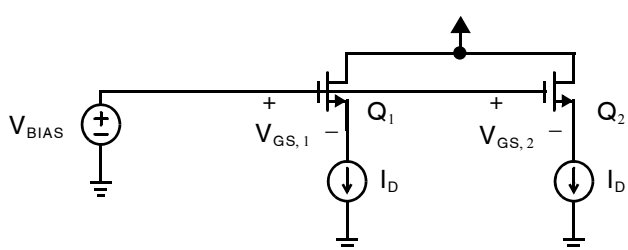


Fig. 2.28 Transistors biased with identical currents.

Shown in Fig. 2.28, transistors biased with the same currents will have gate-source voltages that vary as follows:

$$\sigma^2(\Delta V_{GS}) = \sigma^2(\Delta V_T) + \left(\frac{\sigma(\Delta K')}{K'}\right)^2\left(\frac{I_D}{g_m}\right)^2$$
$$= \frac{1}{WL}\left[A_{Vt0}{}^2\left(\frac{g_m}{I_D}\right)^2 + A_{K'}{}^2\right] \qquad (2.19)$$

where it is assumed in the second line that the devices are closely spaced. Again, V_t-variations are likely to dominate in practice. This may be of interest in knowing how much headroom is available at a particular point in a circuit.

2.4 ANALOG LAYOUT CONSIDERATIONS

When one designs analog circuits, several important layout issues should be considered to realize high-quality circuits. Firstly, the layout should avoid sources of systematic variation which will cause device parameters to deviate from their expected values. Second, good layout practices will minimize noise and outside interference in the circuit.

2.4.1 Transistor Layouts

> ***Key Point:*** *In a common centroid layout, any linear gradient in electrical properties across the chip effects two or more devices equally. This is achieved by dividing each device into smaller equal-sized units, and arranging the units so that all devices have the same mean position ("centroid").*

Transistors in analog circuits are typically much wider than transistors in digital circuits. For this reason, they are commonly laid out using multiple-gate fingers similar to the layout shown in Fig. 2.18. When precision matching between transistors is required, then not only should the individual transistors be realized by combining copies of a single-sized unit transistor, but the fingers for one transistor should be interdigitated with the fingers of the second transistor. Specifically, a *common-centroid layout* is one where the fingers are ordered so that a linear gradient across the integrated circuit in any device parameter, such as the temperature or the gate-oxide thickness, has zero net effect on device performance. This is achieved by ensuring the mean position of the devices are the same. An example of a common-centroid layout for the two identically matched transistors in Fig. 2.27 whose sources are connected is shown in Fig. 2.29 in simplified form [O'Leary, 1991; Maloberti, 1994]. Each of the two transistors

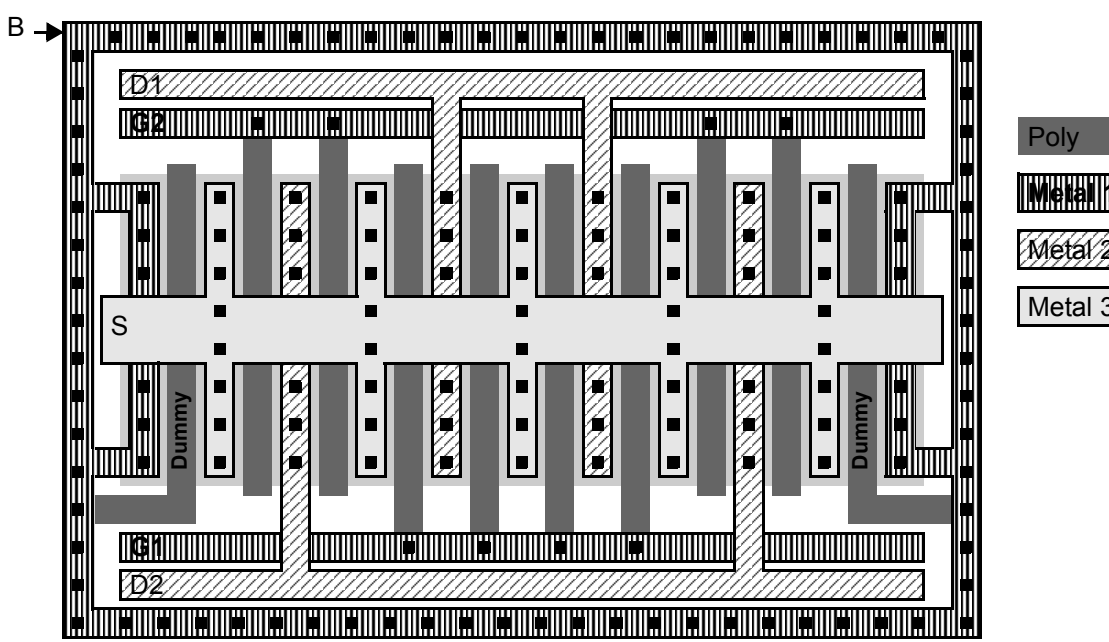


Fig. 2.29 A common-centroid layout for the differential source-coupled pair in Fig. 2.27.

is composed of four separate transistor fingers connected in parallel. The layout is symmetric in both the x and y axes, and any gradients across the microcircuit would affect both M_1 and M_2 in the same way. This layout technique greatly minimizes nonidealities such as opamp input-offset voltage errors when using a differential pair in the input stage of an opamp. Also note that inside the structure, the fingers occur in doubles—two for M_2, two for M_1, two for M_2, and so on. This permits the use of only 2 drain junctions for 4 fingers, reducing the capacitance on that node. The drain connections for each transistor are routed towards the side of the transistor opposite the gate routing. This minimizes parasitic gate-drain (Miller) capacitance, which has an important influence in circuit bandwidth. For the greatest accuracy, only the inside fingers are active. Outside, or dummy, fingers are included only for better matching accuracy and have no other function. The gates of these dummy fingers are normally connected to the most negative power-supply voltage to ensure they are always turned off (or they are connected to the positive power supply in the case of p-channel transistors). A ring of contacts, called a *guard ring*, encircles the transistor OD region providing a low resistance connection to the body terminal under the transistors.

When current mirrors with ratios other than unity are required, again, each of the individual transistors should be realized from a single unit-sized transistor. For example, if a current ratio of 1:2 were desired, then the input transistor might be made from four fingers, whereas the output transistor might be realized using eight identical fingers.

2.4.2 Capacitor Matching

Very often, analog circuits require precise ratios of capacitors. Ideally, capacitor size is given by

$$C_1 = \frac{\varepsilon_{ox}}{t_{ox}} A_1 = C_{ox} x_1 y_1 \tag{2.20}$$

The major sources of errors in realizing capacitors are due to overetching (which causes the area to be smaller than the area of the layout masks) and an oxide-thickness gradient across the surface of the microcircuit. The former effect is usually dominant and can be minimized by realizing larger capacitors from a parallel combination of smaller, unit-sized capacitors, similar to what is usually done for transistors. For example, to realize two capacitors that have a ratio of 4:6, the first capacitor might be realized from four unit-sized capacitors, whereas the second capacitor might be realized by six unit-sized capacitors. Errors due to the gradient of the oxide thickness can then be minimized by interspersing the unit-sized capacitors in a common-centroid layout so the gradient changes affect both capacitors in the same way. Since oxide-thickness variations are not usually large in a reasonably small area, this common-centroid layout is reserved for situations where very accurate capacitors are required.

If only unit-sized capacitors are used, then any overetching will leave the capacitor ratio unaffected. Thus, good designers strive to realize circuits in which only unit-sized capacitors are needed. Unfortunately, this situation is not always possible. When it is not, *overetching error can still be minimized by realizing a nonunit-sized capacitor with a specific perimeter-to-area ratio*. To determine the correct ratio, first note that the error due to overetching is roughly proportional to the perimeter of the capacitor. Specifically, if we assume that a capacitor has an absolute overetching given by Δe and that its ideal dimensions are given by x_1 and y_1, then its true dimensions are given by $x_{1a} = x_1 - 2\Delta e$ and $y_{1a} = y_1 - 2\Delta e$, and the true capacitor size is given by

$$C_a = C_{ox} x_{1a} y_{1a} = C_{ox}(x_1 - 2\Delta e)(y_1 - 2\Delta e) \tag{2.21}$$

This situation is illustrated in Fig. 2.30. Thus, the error in the true capacitance is given by

$$\Delta C_t = C_{ox} x_{1a} y_{1a} - C_{ox} x_1 y_1 = C_{ox}[-2\Delta e(x_1 + y_1) + 4\Delta e^2] \tag{2.22}$$

When this error is small, then the second-order error term can be ignored and (2.22) can be approximated by

$$\Delta C_t \cong -2\Delta e(x_1 + y_1) C_{ox} \tag{2.23}$$

The relative error in the capacitor is therefore given by

$$\varepsilon_r = \frac{\Delta C_t}{C_{ideal}} \cong \frac{-2\Delta e(x_1 + y_1)}{x_1 y_1} \tag{2.24}$$

Thus, the relative capacitor error is approximately proportional to the negative of the ratio of the ideal perimeter to the ideal area (assuming only small errors exist, which is reasonable since, if the errors were not small, then that capacitor sizing would probably not be used). When we realize two capacitors that have different sizes, usually the ratio of one capacitor to the other is important, rather than their absolute sizes. This ratio is given by

$$\frac{C_{1a}}{C_{2a}} = \frac{C_1(1 + \varepsilon_{r1})}{C_2(1 + \varepsilon_{r2})} \tag{2.25}$$

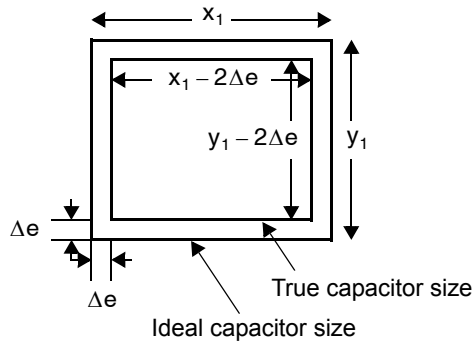


Fig. 2.30 Capacitor errors due to overetching.

If the two capacitors have the same relative errors (i.e., $\varepsilon_{r1} = \varepsilon_{r2}$), then their true ratio is equal to their ideal ratio even when they are not the same sizes. Using (2.24), we see that the relative errors are the same if they both have the same perimeter-to-area ratio. This leads to the following result: *To minimize errors in capacitor ratios due to overetching, their perimeter-to-area ratios should be kept the same, even when the capacitors are different sizes.*

Normally, the unit-sized capacitor will be taken square. *When a non-unit-sized capacitor is required, it is usually set to between one and two times the unit-sized capacitor and is rectangular in shape, so that it has the same number of corners as the unit-sized capacitor.* Defining K to be a desired non-unit-sized capacitor ratio, we have

$$K \equiv \frac{C_2}{C_1} = \frac{A_2}{A_1} = \frac{x_2 y_2}{x_1^2} \tag{2.26}$$

where C_1, A_1, and x_1 represent the capacitance, area, and side-length of a unit-sized capacitor, respectively. Variables C_2, A_2, x_2, and y_2 are similarly defined, except this non-unit-sized capacitor is now rectangular. Equating the ratios of the perimeters-to-areas implies that

$$\frac{P_2}{A_2} = \frac{P_1}{A_1} \tag{2.27}$$

where P_1 and P_2 represent the perimeters of the two capacitors. Rearranging (2.27), we have

$$\frac{P_2}{P_1} = \frac{A_2}{A_1} = K \tag{2.28}$$

which implies that K can also be written as the ratio of perimeters,

$$K = \frac{x_2 + y_2}{2x_1} \tag{2.29}$$

This can be rearranged to become

$$x_2 + y_2 = 2Kx_1 \tag{2.30}$$

Also rearranging (2.26), we have

$$x_2 = \frac{Kx_1^2}{y_2} \tag{2.31}$$

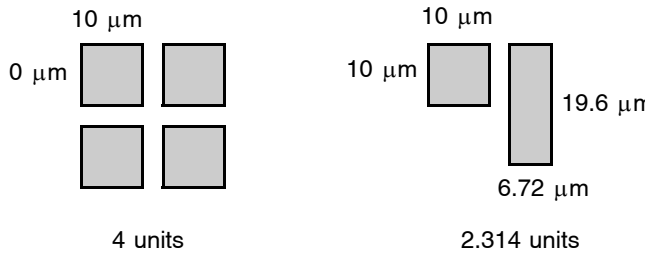


Fig. 2.31 A capacitor layout with equal perimeter-to-area ratios of 4 units and 2.314 units.

Combining (2.30) and (2.31), we find the quadratic equation

$$y_2^2 - 2Kx_1y_2 + Kx_1^2 = 0 \tag{2.32}$$

which can be solved to give

$$y_2 = x_1(K \pm \sqrt{K^2 - K}) \tag{2.33}$$

Recall that K is assumed to be greater than one, which ensures that the square root in (2.33) is applied to a positive number. The value for x_2 is then given by (2.31).

EXAMPLE 2.6

Show a layout that might be used to match two capacitors of size 4 and 2.314 units, where a unit-sized capacitor is 10 μm × 10 μm.

Solution

Four units are simply laid out as four unit-sized capacitors in parallel. We break the 2.314-unit capacitor up into one unit-sized capacitor in parallel with another rectangular capacitor of size 1.314 units. The lengths of the sides for this rectangular capacitor are found from (2.33), resulting in

$$y_2 = 10\,\mu\text{m}\left(1.314 \pm \sqrt{1.314^2 - 1.314}\right) = 19.56\,\mu\text{m or } 6.717\,\mu\text{m}$$

Either of these results can be chosen for y_2, and the other result becomes x_2 ; in other words, the choice of sign affects only the rectangle orientation. Thus, we have the capacitor layout as shown in Fig. 2.31 Note that the ratio of the area of the rectangular capacitor to its perimeter equals 2.5, which is the same as the ratio for the unit-sized capacitor.

Several other considerations should be observed when realizing accurate capacitor ratios. Usually the bottom plate of capacitors is shared by many unit-size capacitors.[5] The interconnection of the top plates can often be done in metal with contacts to the capacitor plates. The parasitic capacitances of any tabs required to contact the plates should be matched as much as possible. This matching often entails adding extra tabs that are not connected anywhere. Another common matching technique is to ensure that the boundary conditions around the unit-sized capacitors match. This boundary-condition matching is accomplished by adding top-plate material around the outside boundaries of unit-sized capacitors at the edge of an array. Many of these principles are illustrated in the simplified layout of two capacitors shown in Fig. 2.32. Each capacitor in the figure consists of two unit-sized capacitors. An additional technique sometimes used (not shown in Fig. 2.32) is to cover the top plates

5. It is usually possible to realize the bottom plate over the top of a well region that has many contacts connected to a low-noise power-supply voltage. This well region acts as a shield to help keep substrate noise out of the bottom plate of the capacitor.

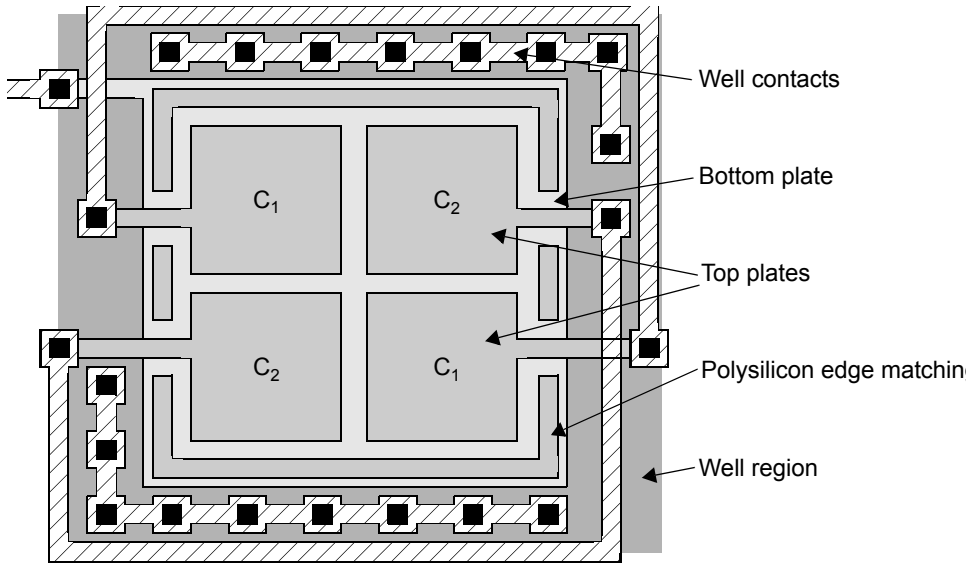


Fig. 2.32 A simplified layout of a capacitor array.

with a layer of metal, which is then connected to the bottom plate. This sandwich-like structure not only gives additional capacitance per area but more importantly, shields the "top" plate from electromagnetic interference. The capacitor plate inside the sandwich is then connected to critical nodes, such as the inputs of amplifiers. Another possibility is to employ interdigitated metal as capacitors, as shown in Fig. 1.37(*b*).

An important consideration when using capacitors in switched-capacitor circuits is that one plate usually has more noise coupled into it than the other because of differing parasitic capacitances to a noisy substrate and/or supply voltage line. Therefore the quieter "top" plate should be connected to critical nodes such as the virtual inputs of opamps, whereas the more noisy "bottom" plate can be connected to less critical nodes such as opamp outputs.

For more details concerning the realization of integrated capacitors, the interested reader can see [Allstot, 1983; O'Leary, 1991; Maloberti, 1994].

2.4.3 Resistor Layout

Integrated resistors can be realized using a wide variety of different conductors. A popular choice is polysilicon, which is a deposited and etched material. Other choices include diffused or ion-implanted regions such as junctions, wells, or base regions. Another possibility is deposited and etched thin-film resistors such as nichrome (consisting of 80 percent nickel and 20 percent chromium) or tantalum. The temperature coefficient of ion-implanted or diffused resistors tends to be positive and large (especially for larger resistivities) with values as large as 1000 to 3000 ppm/°C. On the other hand, the temperature coefficient for thin-film resistors can be as small as 100 ppm/°C. Polysilicon resistors usually have large positive temperature coefficients (say, 1000 ppm/°C) for low-resistivity polysilicon, and moderately large negative temperature coefficients for specially doped high-resistivity polysilicon. The positive temperature coefficients are primarily due to mobility degradation that results from temperature increases. In implanted and diffused resistors, nonlinear resistance varies greatly with voltage because the depletion-region width is dependent on voltage in the more heavily doped conductive region. This depletion-region width variation is substantially smaller in a polysilicon resistor, which is one of the major

reasons polysilicon resistors are preferred over implanted resistors even though they often require more area due to the low resistivity. When thin-film resistors are available in a particular technology, they are almost always the preferred type—unfortunately, they are seldom available.

Regardless of the type of resistor used, the equations governing the resistance (see the Appendix of Chapter 1) are given by

$$R_{\square} = \frac{\rho}{t} \tag{2.34}$$

where $R_{\square}$ is the resistance per square, $\rho = 1/(q\mu_n N_D)$ is the resistivity,[6] t is the thickness of the conductor, and N_D is the concentration of carriers, which we assume are electrons. The total resistance is then given by

$$R = \frac{L}{W} R_{\square} \tag{2.35}$$

where L is the length of the resistor and W is the width of the resistor.

The typical resistivity encountered in integrated circuits depends upon the material employed. For example, in a modern process the gates are formed from a sandwich of a refractory metal over polysilicon (a salicide) with a resistivity of perhaps 1–5 Ω/□, making the gate layer useless for realizing moderate-sized resistors. However, a common way to form resistors inside microcircuits is to use a polysilicon layer without the salicide. This type of polysilicon resistor can have a sheet resistance on the order of 100–1500 Ω/□ depending upon process parameters. This value may vary by 10 – 25% with process and temperature. However, with accurate common centroid layout techniques, very good matching may be obtained between resistors on the same microcircuit.

To obtain large-valued resistors, one must usually use a serpentine layout similar to that shown in Fig. 2.33. Individual resistor fingers are connected in series by metal. The width of each metal is generally

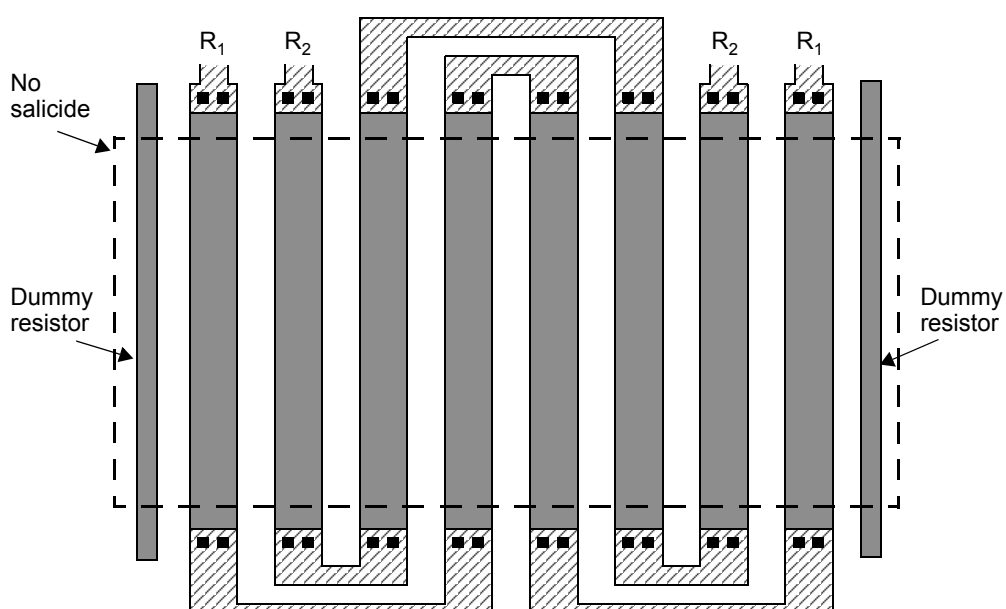


Fig. 2.33 A layout for two matched resistors.

6. This equation is valid for resistors with electron carriers. For resistors with holes as carriers, $\rho = 1/(q\mu_p N_A)$.

far greater than the minimum permitted by the design rules to minimize variations due to edge roughness. Hence, the layout of medium or large-valued resistors can easily be much greater than that of the surrounding transistors. Assuming polysilicon resistors, the metal contacts require a salicide to be deposited on the ends of each strip. In addition to the strips' sheet resistance, additional series resistances appear at the interfaces between salicided to non-salicided polysilicon and due to the contacts. For example, roughly $20\ \Omega$ per contact may be added, but this value is very poorly controlled and may exhibit process variations of over 50%. Hence, the layout should ensure that the contact and interface resistances are a small fraction of the overall resistance. Specifically, very short fingers should be avoided and multiple contacts should always be used in parallel.

The layout in Fig. 2.33 shows two resistors, intended to be well-matched. Two dummy fingers have been included at either side of the layout to match boundary conditions. This structure might result in about 0.1 percent matching accuracy of identical resistors if the finger widths are taken much wider than the fabrication process's minimum feature size. As with integrated capacitors, it is also a good idea to place a shield under a resistor that is connected to a clean power supply. An appropriate shield might be a well region. This shielding helps keep substrate noise from being injected into the conductive layer. (Noise is due to capacitive coupling between the substrate and a large resistor structure.) Also, the parasitic capacitance between the resistor and the shield should be modelled during simulation. Its second-order effects on circuits such as RC filters can often be eliminated using optimization, which is available in many SPICE-like simulators. For low-noise designs, a metal shield over the top of a resistor may also be necessary, although it will result in a corresponding increase in capacitance.

For more information on realizing accurate resistor ratios, the reader is referred to [O'Leary, 1991; Maloberti, 1994].

EXAMPLE 2.7

Estimate the resistance of the layout in Fig. 2.33 assuming a sheet resistance of $800\ \Omega/\square$ for the non-salicided polysilicon and $20\ \Omega$ per contact.

Solution

Each finger in the layout comprises a strip of non-salicided polysilicon $10\ \square$ long, hence having a resistance of $10\ \square \times 800\ \Omega/\square = 8\ \text{k}\Omega$. Each finger also includes two parallel contacts at either end accounting for an additional $2 \times (20\ \Omega/2) = 20\ \Omega$ series resistance. Each resistor comprises 4 fingers in series, and therefore has a total resistance of approximately

$$4 \times (8\ \text{k}\Omega + 20\ \Omega) \cong 32\ \text{k}\Omega$$

2.4.4 Noise Considerations

Some additional layout issues help minimize noise in analog circuits. Most of these issues either attempt to minimize noise from digital circuits coupling into the substrate and analog power supplies, or try to minimize substrate noise that affects analog circuits.

With mixed analog–digital circuits, it is critical that different power-supply connections be used for analog circuits than for digital circuits. Ideally, these duplicate power supplies are connected only off the chip. Where a single I/O pin must be used for the power supply, it is still possible to use two different bonding wires extending from a single-package I/O pin to two separate bonding pads on the integrated circuit. At a very minimum, even if a single

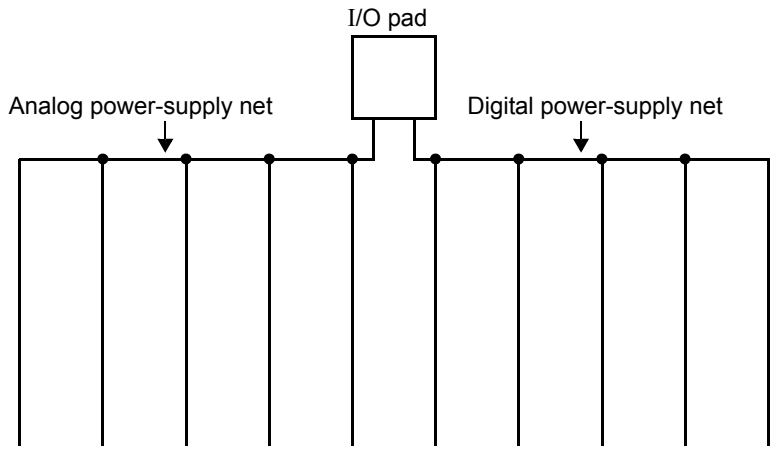


Fig. 2.34 Using separate nets for analog and digital power supplies.

bonding pad is used for both analog and digital circuitry, two separated nets from the bonding pad out should be used for the different types of circuitry, as Fig. 2.34 shows. The reason the power-supply interconnects must be separated is that the interconnect does not have zero impedance. Every time a digital gate or buffer changes state, a glitch is injected on the digital power supply and in the surrounding substrate. By having the analog power supplies separate, we prevent this noise from affecting the analog circuitry. In the ideal case, separate pins are used for the positive power supply and for ground in both the digital and analog circuits. In addition, another pair of pins may be used for the supply voltage and ground for digital output buffers, which can inject very large current spikes. Finally, sometimes multiple pins are used for additional supply and grounds for very large microcircuits.

Another common precaution is to lay out the digital and analog circuitry in different sections of the microcircuit. The two sections should be separated by guard rings and wells connected to the power-supply voltages, as Fig. 2.35 shows. The p^+ connections to ground help keep a low-impedance path between the substrate and

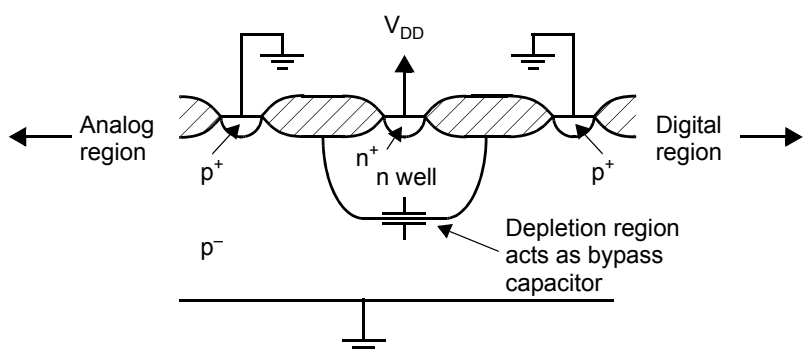


Fig. 2.35 Separating analog and digital areas with guard rings and wells in an attempt to minimize the injection of noise from digital circuits into the substrate under the analog circuit.

ground. For modelling purposes, the substrate can be modelled as a number of series-connected resistors with the p^+ ground connections modelled as resistor-dividers having a small impedance to ground. These low-impedance ground connections help keep substrate noise from propagating through the resistive substrate. The use of the n well between p^+ connections helps to further increase the resistive impedance of the substrate between the analog and digital regions due to graded substrate doping. Specifically, the p^- substrate often has 10 times higher doping at the surface of the microcircuit compared to the doping level below the n well, which leads to a tenfold increase in substrate resistivity between the two p^+ connections. Finally, the n well also operates as a bypass capacitor to help lower the noise on V_{DD}.

Another important consideration when laying out a circuit that includes both analog and digital circuits is the use of shields connected to either ground or to a separate power-supply voltage. Figure 2.36 shows examples of the use of shields. In this example, an n well is used to shield the substrate from the digital interconnect line. The well is also used to shield an analog interconnect line from any substrate noise. This shield is ideally connected to a ground net that is used only for shields. If this type of connection is not possible due to layout and space constraints, then the digital ground can be used for the shields, although this is not ideal. In the example, the shield ground is also connected to metal lines that separate the analog and digital lines from each other and from other interconnect lines. Finally, an additional metal shield might be placed above the lines as well. This final shield may be somewhat excessive, but it can often be easily realized in many parts of the microcircuit if ground and power-supply lines are distributed in metal 2, perpendicular to the metal-1 interconnect lines. It should also be mentioned that the n well shield also acts as a bypass capacitor; this helps minimize noise in the substrate, which is connected to V_{DD}. Additional layers that are often used as shields are the polysilicon layers.

Perhaps the most effective technique for minimizing the propagation of substrate noise in a mixed-mode microcircuit containing both analog and digital circuitry is the use of an epitaxial process. An epitaxial process places a conductive layer under all transistors. Any charge flowing through the substrate is attracted to this layer and does not propagate into sensitive analogs regions. Although this process is more expensive, it also helps prevent latch-up. For deep submicron technologies, this epitaxial process is common because of its reduced latch-up sensitivity.

Careful thought should go into the overall placement of different blocks in a mixed-mode analog-digital microcircuit. A possible arrangement for an analog section containing switched-capacitor circuits is shown in Fig. 2.37. Notice that an n well shield is being used under the capacitors. Notice also that the clock lines are not only as far from the opamps as possible, but are also separated by two wells and a V_{SS} interconnect that is liberally connected to the substrate, two ground (Gnd) lines, and a V_{DD} line. A well is placed under the clock lines as a shield. This shield is connected to a

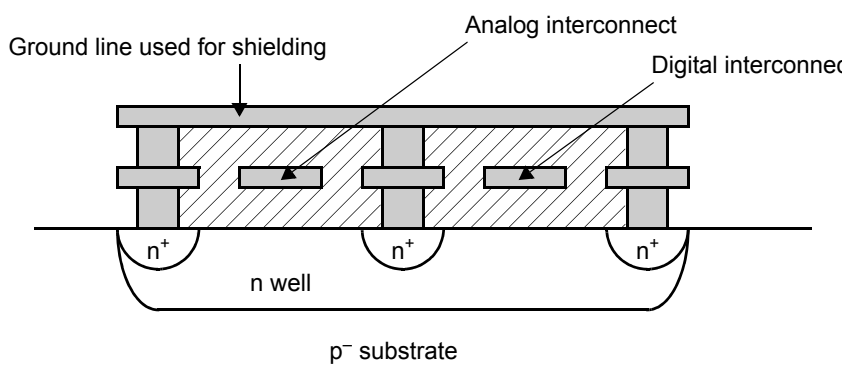


Fig. 2.36 Using shields helps keep noise from being capacitively coupled into and out of the substrate.

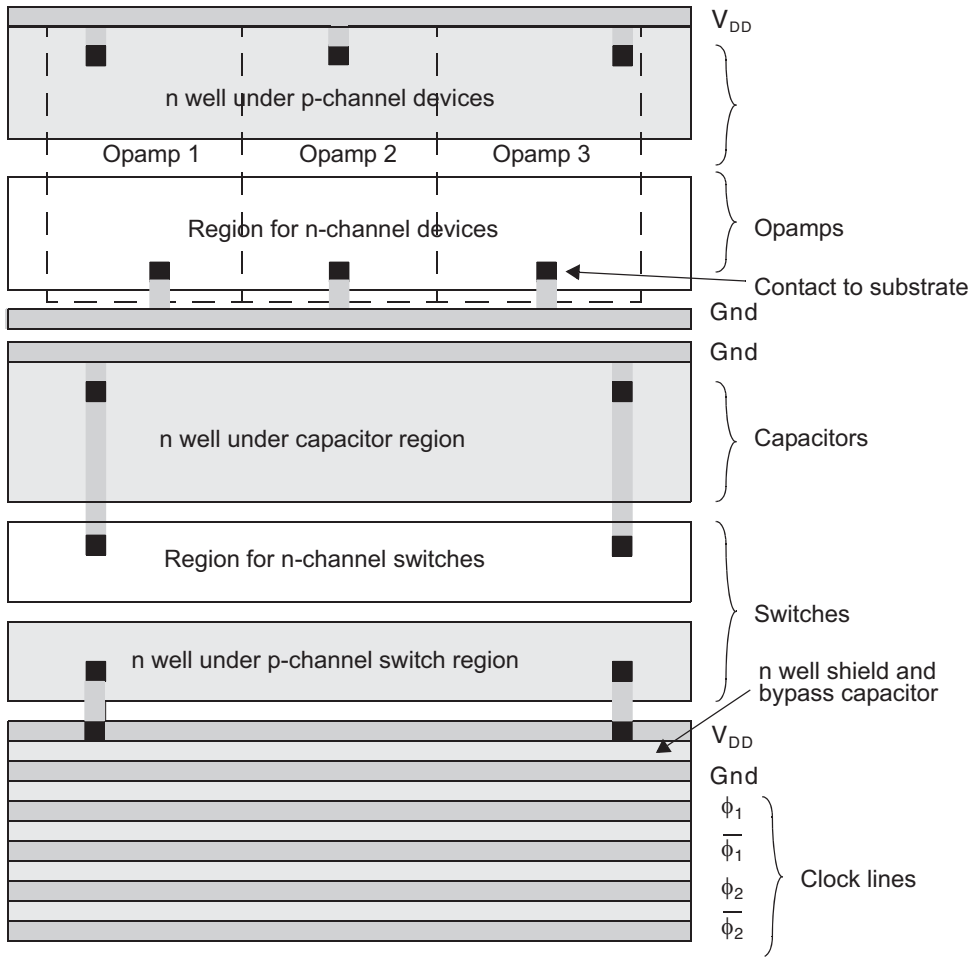


Fig. 2.37 A possible floor plan for an analog section containing switched-capacitor circuits.

separate ground line (perhaps digital ground) from the one used in the opamp region because this shield will likely have quite a bit of clock noise coupled into it. Also note that a separate V_{DD} line is used to connect to the n wells under the switches, a region where digital interconnects exist, as is used in the critical opamp section.

One last technique for noise minimization in analog microcircuits should always be used: After layout has been finished, *any unused space should be filled with additional contacts to both the substrate and to the wells, which are used as bypass capacitors.* In a typical microcircuit, this results in a significant increase in bypass capacitance.

Many other techniques have been developed by various companies, but the preceding techniques give the reader a good idea of the types of practical considerations necessary when realizing high-performance analog microcircuits.

2.5 KEY POINTS

- The first step in realizing an integrated circuit is to produce a single-crystalline silicon wafer from 10 to 30 cm in diameter and roughly 1 mm thick. The silicon is either lightly doped, or heavily doped with a thin lightly-doped epitaxial layer on top in which the transistors are made. [p. 74]
- Most features on integrated circuits are patterned using photolithography whereby light is passed through a mask to cast patterns onto the underlying silicon wafer, ultimately defining the circuit's physical features such as transistor sizes and wiring. [p. 75]
- Transistors are fabricated inside the "active" or "oxide definition" (OD) regions of the microcircuit. Over OD regions, only a very thin oxide separates the polysilicon gates from the transistor channel regions underneath, and additional dopants are introduced to control the threshold voltages. Surrounding the OD regions are isolation structures to prevent parasitic transistors from conducting leakage currents. [p. 81]
- The edge of transistor source and drain junctions are defined by the edge of the polysilicon gate above. This "self-alignment" of the gate and junctions was key in the development of small high-speed transistors. [p. 83]
- Up to 10 or even more layers of metal are patterned above the silicon surface, separated by insulating oxide, to provide interconnect between all the devices in a circuit. [p. 84]
- Finite tolerances during integrated circuit manufacturing impose constraints on the minimum sizes and spacing of transistor and interconnect features. These constraints may be expressed as multiples of λ, equal to one-half the minimum gate length. They influence parasitic capacitances, and ultimately the circuit bandwidth which an analog designer may expect. [p. 88]
- Systematic variations are those observed repeatedly and consistently, even when a circuit is being mass-produced. Generally, these may be avoided by proper layout techniques, although doing so may impose some penalty in terms of layout size or performance. Process variations are observed because the conditions of manufacture (temperature, concentration levels, etc.) can never be maintained precisely constant. These manifest as device parameters that differ from one sample of a circuit to the next. Random variations are statistical in nature and are present in every individual device. Hence, they may be observed as mismatch between two identically specified transistors fabricated under the same nominal conditions. [p. 96]
- The absolute sizes and electrical parameters of integrated circuit components can seldom be accurately determined. For best accuracy, larger objects are made out of several unit-sized components connected together, and the boundary conditions around all objects should be matched, even when this means adding extra unused components. [p. 98]
- Practical variations in transistor parameters, supply voltage, and temperature (PVT variations) can be a bane to analog designers trying to reliably meet specific performance criteria. [p. 99]
- The parameters of devices that are closely spaced with identical layout exhibit a random variance inversely proportional to device area. In practice, the most important random variations in most analog circuits are mismatch. [p. 100]
- In a common centroid layout, any linear gradient in electrical properties across the chip effects two or more devices equally. This is achieved by dividing each device into smaller equal-sized units, and arranging the units so that all devices have the same mean position ("centroid"). [p. 103]

2.6 REFERENCES

D. Allstot and W. Black. "Technological Design Considerations for Monolithic MOS Switched-Capacitor Filtering Systems," *IEEE Proceedings*, Vol. 71, no. 8, pp. 967–986, August 1983.

P. G. Drennan, M. L. Kniffin, and D. R. Locascio. "Implications of Proximity Effects for Analog Design," in *Custom Integrated Circuits Conference*, Sept. 2006.

R. Geiger, P. Allen, and N. Strader. *VLSI: Design Techniques for Analog and Digital Circuits*. McGraw-Hill, New York, 1990.

A. Glasser and G. Subak-Sharpe. *Integrated Circuit Engineering: Design, Fabrication, and Applications*. Addison-Wesley, Reading, Massachusetts, 1977.

A. Grebene, *Bipolar and MOS Analog Integrated Circuits*. Wiley-Interscience, Wiley, New York, 1984.

R. Haveman, et al. "A 0.8-μm, 256K BiCMOS SRAM Technology," *Digest of Technical Papers, 1987 Intern. Electron Devices Meeting*, pp. 841–843, December 1987.

R. Haken, R. Haveman, R. Ekund, and L. Hutter. "BiCMOS Process Design," in *BiCMOS Technology and Applications*, ed. A. Alvarez. Kluwer Academic Publishers, Nowell, Massachusetts, 1989.

P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE J. Soild State Circuits*, vol. 40, no. 6, pp. 1212–1224, June 2005.

F. Maloberti. "Layout of Analog and Mixed Analog-Digital Circuits," in *Design of Analog-Digital VLSI Circuits for Telecommunication and Signal Processing*, ed. J. Franca and Y. Tsividis. Prentice Hall, Englewood Cliffs, New Jersey, 1994.

S. Muraka and M. Peckerar. *Electronic Materials.* Academic Press, New York, 1989, p. 326.

P. O'Leary. "Practical Aspects of Mixed Analogue and Digital Design," in *Analogue-Digital Asics, Circuit Techniques, Design Tools, and Applications*, ed. R. S. Soin, F. Maloberti, and J. Franca. Peter Peregrinus, Stevenage, England, 1991.

M. Pelgrom, A. Duinjamer, and A. Welbers, "Matching properties of MOS transistors," *IEEE J. Soild State Circuits*, Vol. 24, no. 5, pp. 1433–1439, May 1989.

D. Reinhard. *Introduction to Integrated Circuit Engineering*. Houghton Mifflin, Boston, 1987.

Y. Tsividis, *Mixed Analog-Digital VLSI Devices and Technology (An Introduction)*, World Scientific, Hackensack, New Jersey, 2002.

2.7 PROBLEMS

2.7.1 Section 2.1: CMOS Processing

2.1 Discuss briefly the relationships between an ion beam's acceleration potential, the beam current, and the time of implantation on the resulting doping profile.

2.2 Place the following processing steps in their correct order: metal deposition and patterning, field implantation, junction implantation, well implantation, polysilicon deposition and patterning, field-oxide growth.

2.3 What are the major problems associated with a single threshold-voltage-adjust implant?

2.4 What is the reason for using a field implant and why is it often not needed in the well regions?

2.5 What are the major trade-offs in using a wet process or a dry process when growing thermal SiO_2?

2.6 Why is polysilicon rather than metal used to realize gates of MOS transistors?

2.7 Why can't a microcircuit be annealed after metal has been deposited?

2.7.2 Section 2.2: CMOS Layout and Design Rules

2.8 What minimum distance, in terms of λ, would you expect that metal should be separated from polysilicon? Why?

2.9 Find the circuit that the layout shown in Fig. P2.9 realizes. Simplify the circuit, if possible, and give the sizes of all transistors. Assume $L = 2\lambda$, where $\lambda = 0.2\ \mu m$.

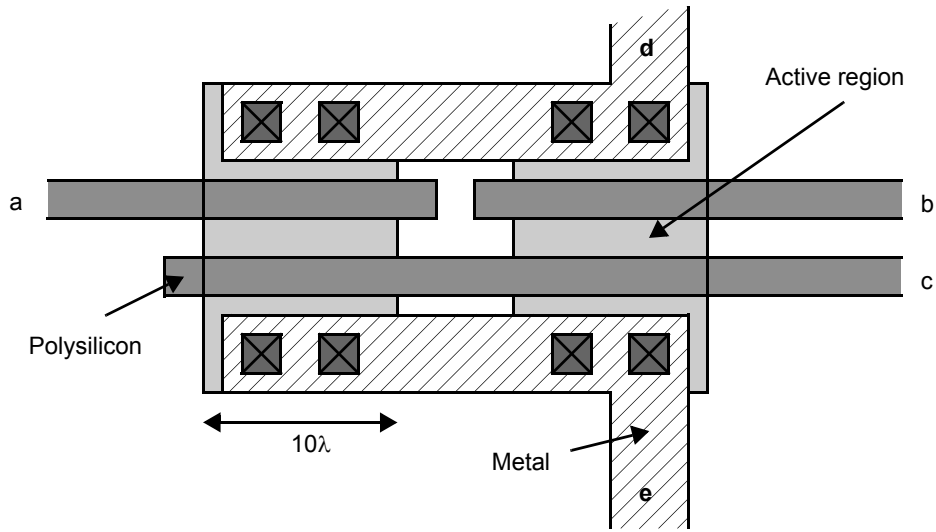


Fig. P2.9

2.10 Find the transistor schematic for the CMOS logic circuit realized by the layout shown in Fig. P2.10. Give the widths of all transistors. Assume L = 2λ, where λ = 0.4 μm. In tabular form, give the area and perimeter of each junction that is not connected to VDD or to ground.

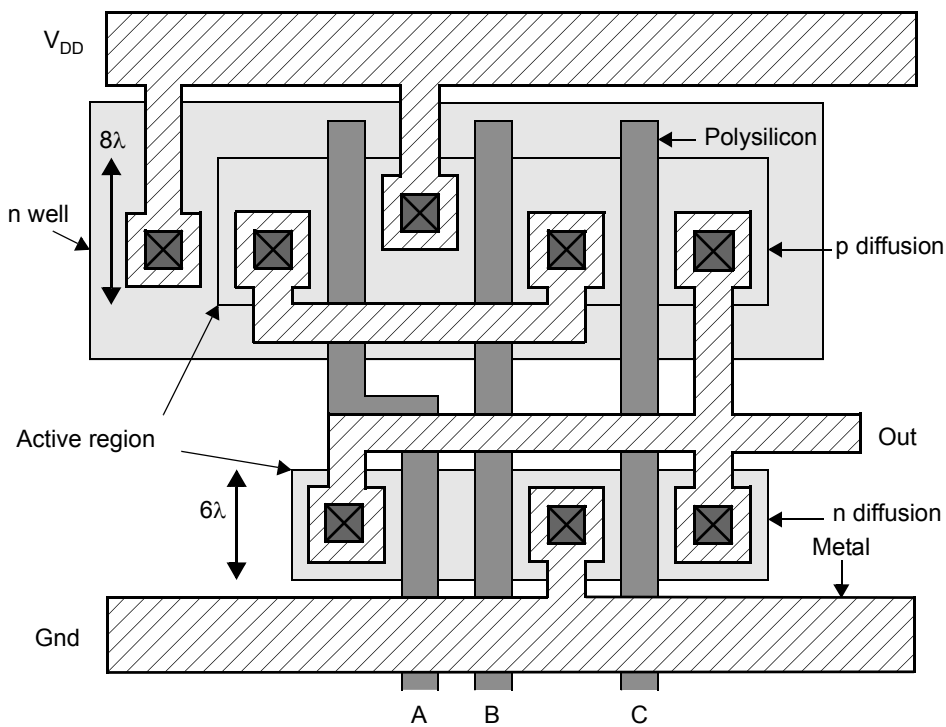


Fig. P2.10

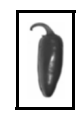
2.11 Repeat Example 2.1 for the case in which the two transistors do not physically share any junction, but each junction is realized in a way similar to junction J_2.

2.12 Repeat Example 2.2 where an overall transistor width of 80λ is still desired, but assume 8 parallel transistors, each of width 10λ, are used.

2.13 Repeat Example 2.2 where an overall transistor width of 80λ is still desired, but assume 2 parallel transistors, each of width 40λ, are used.

2.7.3 Section 2.3: Variability and Mismatch

In the following problems, use the device parameters listed for the 0.18-μm CMOS process in Table 1.5. Assume that $A_{Vt0} = 4\ \text{mV} \cdot \mu\text{m}$ and $A_{K'} = 0.01\ \mu\text{m}$.



2.14 For a NMOS transistor sized $(W/L) = (20\ \mu\text{m}/0.2\ \mu\text{m})$ and biased with a fixed $I_D = 0.5$ mA, how much does the unity-gain frequency of the transistor, f_T, change with a 5% decrease in C_{ox} and a 10% decrease in μ_n?

2.15 An NMOS triode device is biased with a fixed $V_{GS} = 1$ V. Size the device so that it has a small signal resistance of 300 Ω or less under 3σ statistical variations.

2.16 Repeat problem 2.15 for a PMOS device with a fixed $V_{GS} = 1$ V.

2.17 Size the devices in a simple PMOS current mirror to provide a nominal drain current of 100 μA with a small signal output resistance of at least 5 kΩ and to have 3σ random variations result in 1% change their drain current.

2.18 Size NMOS differential pair devices to have a transconductance of 1 mA/V with a tail current of 500 μA and an input offset better than 2 mV 99.8% of the time.

2.19 For the simple actively loaded common source amplifier shown in Fig. P2.19, find an expression for the contribution of Q_2/Q_3 mismatch towards variations in the transconductance of Q_1. Assume that the dc bias voltage on V_{in} is somehow automatically varied to keep all transistors in active mode.

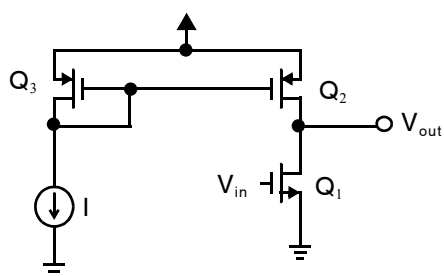


Fig. P2.19

2.7.4 Section 2.4: Analog Layout Considerations

2.20 We desire to match two capacitors of relative sizes 9 and 4.523. Sketch a layout for the two capacitors such that their ratio will be maintained during overetching.

2.21 Given that a polysilicon layer has 1 kΩ/□, what is the resistance of a long line that is 2 μm wide and 100 μm long? (Ignore any contact resistance.) Sketch a layout made of polysilicon strips, each 10□ long.

CHAPTER

3 Basic Current Mirrors and Single-Stage Amplifiers

In this chapter, fundamental building blocks are described. These blocks include a variety of current mirrors, single-stage amplifiers with active loads, and differential pairs. A good knowledge of these building blocks is critical to understanding many subjects in the rest of this book and for analog IC design in general. CMOS mirrors and gain stages are emphasized because they are prevalent in modern designs. Fortunately, most of the small-signal analyses presented can be applied to bipolar circuits with little change. In addition, rather than using resistive loads and ac coupling, the gain stages covered are shown with current-mirror active loads since such loads are almost always used in integrated circuits.

When analyzing electronic circuits containing transistors to determine their small-signal behavior, it is implicitly assumed that signals are small enough that linear approximations about an operating point accurately reflect how the circuit operates. These linear approximations may be represented schematically by replacing transistors with their small-signal equivalents, whose parameters (g_m, r_{ds}, etc.) are related to the device's operating point currents and voltages and are summarized in Section 1.3. The general procedure for small-signal analysis is therefore:

a. Set all signal sources to zero and perform an operating point analysis for all currents and voltages. A voltage source set to 0 V is the same as an ideal wire—a short circuit. A current source set to 0 A is the same as an open circuit.

b. Replace all transistors with their small-signal equivalents where the parameters g_m, r_{ds}, etc, are found from the operating point voltages and currents using the relationships in summarized in Section 1.3.

c. Set all independent sources equal to zero, except for the signal sources that were zeroed in step (a). This includes power supply voltages, bias currents, etc. Remember that setting a voltage source to zero means replacing it with a short circuit, and setting a current source to zero means replacing it with an open circuit.

d. Analyze the resulting linearized small-signal circuit to find small-signal node voltages, branch currents, small-signal resistances, etc.

e. If desired, the complete solution may be found by superimposing the results of the operating point analysis in step (a) and those of small-signal analysis in step (d). The result so obtained is approximate because the small-signal analysis approximates transistor nonlinear behavior with linearized models.

To the extent possible in this chapter, operating point quantities are represented with uppercase voltage and current symbols (e.g., V_{GS}, I_D) and small-signal quantities with lowercase symbols (e.g., v_{gs}, i_d). However, the practicing designer must always be alert to imprecise notation and remain able to interpret meanings within their proper context.

3.1 SIMPLE CMOS CURRENT MIRROR

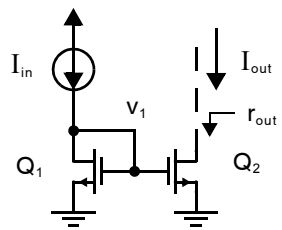


Fig. 3.1 A simple CMOS current mirror.

An ideal current mirror is a two-port circuit that accepts an input current I_{in} and produces and output current $I_{out} = I_{in}$. Since current sensing is best done with a low resistance, as in for example an ammeter, the ideal current source will have zero input resistance. An ideal current source has a high output resistance and, hence, so will an ideal current mirror. In this way, the ideal current mirror faithfully reproduces the input current regardless of the source and load impedances to which it is connected.

A simple CMOS current mirror is shown in Fig. 3.1, in which it is assumed that both transistors are in the active region, which means that the drain voltage of Q_2 must be greater than V_{eff2}. If the finite small-signal drain-source impedances of the transistors are ignored, and it is assumed that both transistors are the same size, then Q_1 and Q_2 will have the same current since they both have the same gate-source voltage, V_{gs}. However, when finite drain-source impedance is considered, whichever transistor has a larger drain-source voltage will also have a larger current. Let us compare this basic circuit to the "ideal" current source with zero input resistance and infinite output resistance.

To find the input resistance, consider the small-signal model for Q_1 alone, as shown in Fig. 3.2(*a*). The independent current source I_{in} does not exist in the small-signal model and is replaced with an open circuit. Also note that a low-frequency small-signal model is used for Q_1 (i.e., all the capacitors are ignored in the model). This small-signal model can be further reduced by finding the *Thévenin-equivalent* circuit. The Thévenin-equivalent output voltage is 0 since the circuit is stable and contains no input signal. This circuit's Thévenin-equivalent input impedance is found by applying a test signal voltage, v_y, at v_1 and measuring the signal current, i_y, as shown. Here, the current i_y is given by

$$i_y = \frac{v_y}{r_{ds1}} + g_{m1}v_{gs1} = \frac{v_y}{r_{ds1}} + g_{m1}v_y \tag{3.1}$$

The input impedance is given by v_y/i_y which equals $1/g_{m1} \parallel r_{ds1}$. Because typically $r_{ds1} >> 1/g_{m1}$, we approximate the input impedance to be simply $1/g_{m1}$ (which is also defined to be r_{s1}) resulting in the equivalent model shown in Fig. 3.2(*c*). This same result holds in the bipolar case and is also equivalent to the small-signal model for a diode. Hence, Q_1 is often referred to as a *diode-connected* transistor.

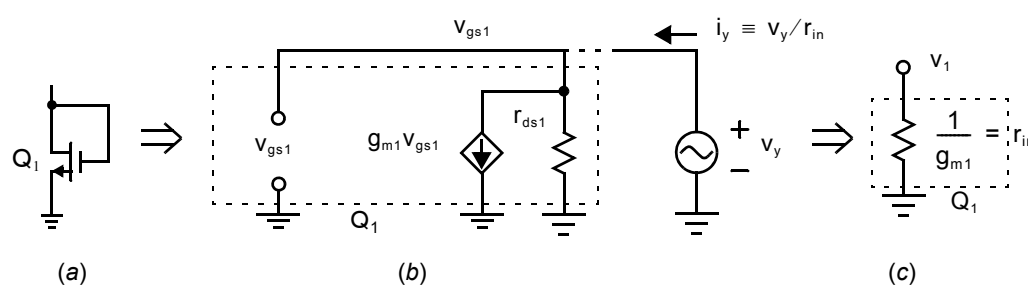


Fig. 3.2 (*a*) A diode-connected transistor, Q_1, (*b*) the small-signal model for Q_1, and (*c*) an equivalent simplified small-signal model for Q_1.

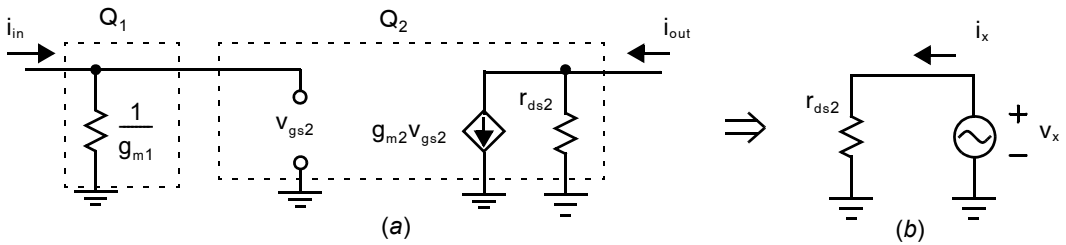


Fig. 3.3 (*a*) A small-signal model for the current mirror of Fig. 3.1 and (*b*) a simplified small-signal model for determining the small-signal output resistance, r_{out}.

Using the model just described leads to a simplified small-signal model for the overall current mirror shown in Fig. 3.3(*a*) where v_{gs2} has been connected to ground via a resistance of $1/g_{m1}$. Since no current flows through the $1/g_{m1}$ resistor, $v_{gs2} = 0$ no matter what voltage v_x is applied to the current-mirror output. This should come as no surprise, since MOS transistors operate unilaterally at low frequencies. Thus, since $g_{m2}v_{gs2} = 0$, the circuit is simplified to the equivalent small-signal model shown in Fig. 3.3(*b*). The small-signal output resistance, r_{out}, is simply equal to r_{ds2}.

> ***Key Point:*** *A simple CMOS current mirror has a small-signal input resistance of* $1/g_{m1}$ *and a small-signal output resistance* r_{ds2}.

EXAMPLE 3.1

Consider the current mirror shown in Fig. 3.1, where $I_{in} = 100\ \mu A$ and each transistor has $W/L = 10\ \mu m/0.4\ \mu m$. Given the 0.35-μm CMOS device parameters in Table 1.5, find r_{out} for the current mirror and the value of g_{m1}. Also, estimate the change in I_{out} for a 100 mV change in the output voltage. What voltage must be maintained at the drain of Q_2 to ensure it remains in active mode?

Solution

Since the W/L ratios of Q_1 and Q_2 are the same, the nominal value of I_{out} equals that of $I_{in} = 100\ \mu A$. Thus, we have

$$r_{out} = r_{ds2} = \frac{L}{\lambda L I_D} = \frac{0.4\ \mu m}{(0.16\ \mu m/V)(100\ \mu A)} = 25\ k\Omega \tag{3.2}$$

The value of g_{m1} is given by

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)I_{D1}} = 0.97\ mA/V \tag{3.3}$$

resulting in $r_{s1} = 1/g_{m1} \cong 1.03\ k\Omega$. Note that this r_{s1} value is significantly less than r_{ds1}, which equals r_{ds2} in this case, so that we may assume $r_{in} \cong 1.03\ k\Omega$.

The change in output current can be estimated, using r_{out}, as

$$\Delta I_{out} = \frac{\Delta V}{r_{out}} = \frac{100\ mV}{25\ k\Omega} = 4\ \mu A \tag{3.4}$$

In other words, if initially I_{out} is measured to be 101 μA (due to mismatch or a larger V_{DS} voltage), then a 100 mV increase in output voltage would result in a new output current of about 105 μA. Note that this estimate does not account for second-order effects such as the fact that r_{ds} changes as the output current changes.

Finally, the drain voltage of Q_2 must be at least V_{eff2} in order to keep it in active mode.

$$V_{eff2} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} = 205 \text{ mV}$$

You may compare these results to those obtained with a SPICE simulation.

3.2 COMMON-SOURCE AMPLIFIER

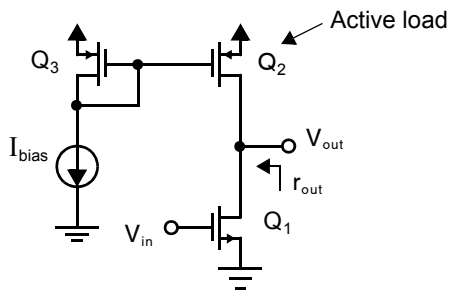


Fig. 3.4 A common-source amplifier with a current-mirror active load.

A common use of simple current mirrors is in a single-stage amplifier with an active load, as shown in Fig. 3.4. This common-source topology is the most popular gain stage, especially when high input impedance is desired.

Here, an n-channel common-source amplifier has a p-channel current mirror used as an active load to supply the bias current for the drive transistor. By using an active load, a high-impedance output load can be realized without using excessively large resistors or a large power-supply voltage. As a result, for a given power-supply voltage, a larger voltage gain can be achieved using an active load than would be possible if a resistor were used for the load. For example, if a 100-kΩ load were required with a 100-μA bias current, a resistive-load approach would require a power-supply voltage of $100\ \text{k}\Omega \times 100\ \mu\text{A} = 10\ \text{V}$. An active load takes advantage of the nonlinear, large-signal transistor current–voltage relationship to provide large small-signal resistances without large dc voltage drops.

Key Point: *The common-source amplifier is a popular gain stage, especially when high input impedance is desired. The use of an active load takes advantage of the nonlinear, large-signal transistor current–voltage relationship to provide large small-signal resistances without large dc voltage drops.*

A small-signal equivalent circuit for the low-frequency analysis of the common-source amplifier of Fig. 3.4 is shown in Fig. 3.5 where V_{in} and R_{in} are the Thévenin equivalent of the input source. It is assumed that the bias voltages are such that both transistors are in the active region. The output resistance, R_2, is made up of the parallel combination of the drain-to-source resistance of Q_1, that is, r_{ds1}, and the drain-to-source resistance of Q_2, that is, r_{ds2}. Notice that the voltage-controlled current source modelling the body effect has not been included since the source is at a small-signal ground, and, therefore, this source always has 0 current.

Using small-signal analysis, we have $v_{gs1} = v_{in}$ and, therefore,

$$A_V = \frac{v_{out}}{v_{in}} = -g_{m1}R_2 = -g_{m1}(r_{ds1} \parallel r_{ds2}) \tag{3.5}$$

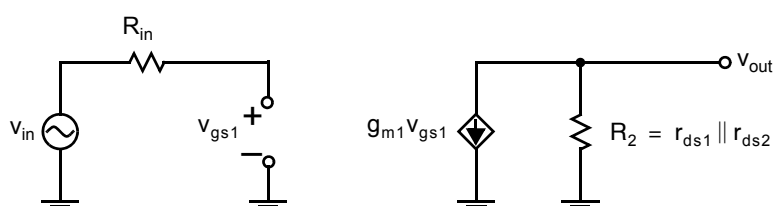


Fig. 3.5 A small-signal equivalent circuit for the common-source amplifier.

Depending on the device sizes, currents, and the technology used, a typical gain for this circuit is in the range of –5 to –100. To achieve similar gains with resistive loads, much larger power-supply voltages must be used which also greatly increases the power dissipation. However, it should be mentioned here that for low-gain, high-frequency stages, it may be desirable to use resistor loads (if they do not require much silicon area) because they often have less parasitic capacitances associated with them. They are also typically less noisy than active loads.

EXAMPLE 3.2

Assume I_{bias} = 100 μA, all transistors have W/L = 10 μm/0.4 μm in Fig. 3.4, and the device parameters are those of the 0.35-μm CMOS process in Table 1.5. What is the gain of the stage?

Solution

We have

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)_1 I_{bias}} = 0.97 \text{ mA/V} \tag{3.6}$$

In this case, λL is the same for both NMOS and PMOS devices, so

$$r_{ds1} = r_{ds2} = \frac{L}{\lambda L I_D} = \frac{0.4 \ \mu\text{m}}{(0.16 \ \mu\text{m/V})(100 \ \mu\text{A})} = 25 \text{ k}\Omega \tag{3.7}$$

Using Eq. (3.5), we have

$$A_V = -g_{m1}(r_{ds1} \parallel r_{ds2}) = -0.97 \text{ mA/V}(25 \text{ k}\Omega \parallel 25 \text{ k}\Omega) = -12.1 \text{ V/V} \tag{3.8}$$

Compare these results to those obtained with a SPICE simulation.

Under the simple assumption that $r_{ds2} \approx r_{ds1}$, the gain of the common source amplifier is one-half the intrinsic gain of transistor Q_1: $A_i = g_{m1}r_{ds1} \approx 2/(\lambda V_{eff})$. Hence, in order to maximize the gain of this stage, it is desirable to maximize the intrinsic gain by operating Q_1 with small V_{eff1}. For a fixed bias drain current, I_D, the effective overdrive voltage is reduced by increasing the device width W. However, beyond a certain width, V_{eff} approaches zero, the transistor will enter subthreshold operation, and no further increases in intrinsic gain are observed.

EXAMPLE 3.3

Modify the design in Example 3.2 to increase the gain by 20% by changing only the device width, W.

Solution

Neglecting higher-order effects, with the drain current fixed the output resistance ($r_{ds1} \parallel r_{ds2}$) is nominally unchanged. Hence, in order to increase gain by 20%, we must increase g_{m1} by 20%. Due to the square-root dependence of g_m on W (with fixed current), this in turn requires a 44% increase in W_1. Hence, the new size of Q_1 is (14.4 μm/0.4 μm). The resulting gain is

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)_1 I_{bias}} = 1.17 \text{ mA/V}$$

$$\Rightarrow A_V = -g_{m1}(r_{ds1} \parallel r_{ds2}) = -1.17 \text{ mA/V}(25 \text{ k}\Omega \parallel 25 \text{ k}\Omega) = -14.6 \text{ V/V}$$

This is, indeed, 20% greater than the gain of –12.1 V/V computed in Example 3.2. The resulting effective gate-source voltage is

$$V_{eff1} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} = 171 \text{ mV}$$

which is sufficient to keep Q_1 out of subthreshold operation.

3.3 SOURCE-FOLLOWER OR COMMON-DRAIN AMPLIFIER

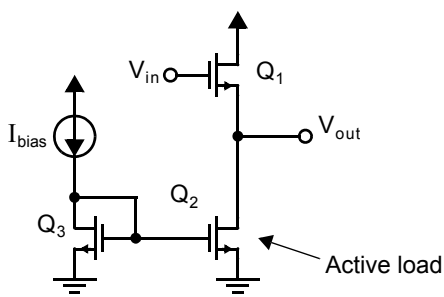


Fig. 3.6 A source-follower stage with a current mirror used to supply the bias current.

Another general use of current mirrors is to supply the bias current of source-follower amplifiers, as shown in Fig. 3.6. In this example, Q_1 is the source follower and Q_2 is an active load that supplies the bias current of Q_1. These amplifiers are commonly used as *voltage buffers* and are therefore commonly called source followers. They are also referred to as common-drain amplifiers, since the input and output nodes are at the gate and source nodes, respectively, with the drain node being at small-signal ground. Although the dc level of the output voltage is not the same as the dc level of the input voltage, ideally the small-signal voltage gain is close to unity. In reality, it is somewhat less than unity. However, although this circuit does not generate voltage gain, it does have the ability to generate current gain.

Key Point: *The source-follower provides a voltage gain close to unity, and often limited by the body effect. It can provide a large current gain and it is unilateral so it is often used as a voltage buffer.*

A small-signal model for low-frequency analysis of this source-follower stage is shown in Fig. 3.7. Note that the voltage-controlled current source that models the body effect of MOS transistors has been included because the source is not at small-signal ground. The body effect is a major limitation on the small-signal gain. Note that in Fig. 3.7, r_{ds1} is in parallel with r_{ds2}. Notice also that the voltage-controlled current source modelling the body effect produces a current that is proportional to the voltage across it. This relationship makes the body effect equivalent to a resistor of size $1/g_{s1}$, which is also in parallel with r_{ds1} and r_{ds2}. Thus, the small-signal model of Fig. 3.7 is equivalent to the simplified small-signal model of Fig. 3.8, in which $R_{s1} = r_{ds1} \parallel r_{ds2} \parallel 1/g_{s1}$. Writing the nodal equation at v_{out}, and noting that $v_{gs1} = v_{in} - v_{out}$, we have

$$v_{out}/R_{s1} - g_{m1}(v_{in} - v_{out}) = 0 \tag{3.9}$$

To minimize circuit equation errors, a consistent methodology should be maintained when writing nodal equations. The methodology employed here is as follows: The first term is always the node at which the currents are being summed. This node voltage is multiplied by the sum of all admittances connected to the node. The next negative terms are the adjacent node voltages, and each is multiplied by the connecting admittance. The last terms are any current sources with a multiplying negative sign used if the current is shown to flow into the node.

Solving for v_{out}/v_{in}, we have

$$A_v = \frac{v_{out}}{v_{in}} = \frac{g_{m1}}{g_{m1} + G_{s1}} = \frac{g_{m1}}{g_{m1} + g_{s1} + g_{ds1} + g_{ds2}} = g_{m1}\left(\frac{1}{g_{m1}} \parallel \frac{1}{g_{s1}} \parallel r_{ds1} \parallel r_{ds2}\right) \tag{3.10}$$

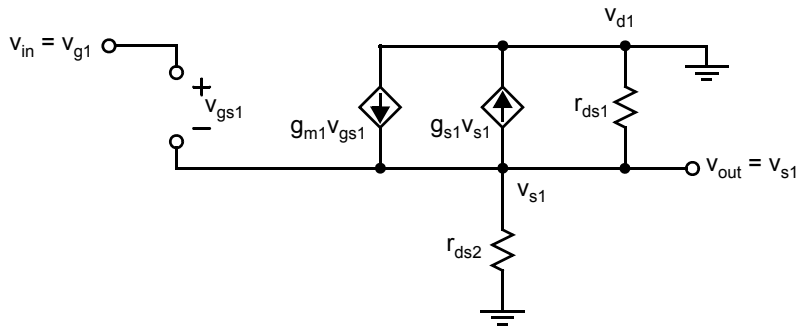


Fig. 3.7 The low-frequency model of the source-follower amplifier.

where the notation $G_{s1} = 1/R_{s1}$ is used.[1] Normally, g_{s1} is on the order of one-tenth to one-fifth that of g_{m1}. Also, the transistor output admittances, g_{ds1} and g_{ds2}, might be one-tenth that of the body-effect parameter, g_{s1}. Therefore, it is seen that the body-effect parameter is the major source of error causing the gain to be less than unity. Notice also that at low frequencies the stage is completely unilateral. In other words, there is no signal flow from the output to the input. This can be seen by applying a small test signal to the output and noting that it induces no voltage or current at the input.

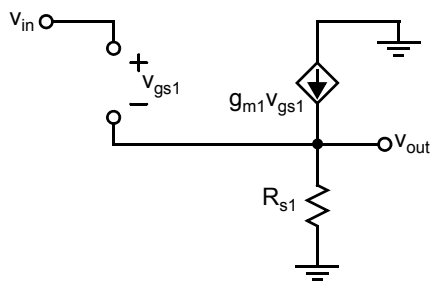


Fig. 3.8 An equivalent small-signal model for the source follower.

EXAMPLE 3.4

Consider the source follower of Fig. 3.6 where $I_{bias} = 100\ \mu A$, all transistors have $W/L = 2\ \mu m/0.2\ \mu m$, $\phi_F \cong 0.4$ V and $\gamma = 0.3\ V^{-1/2}$, and the other device parameters are those of the 0.18-μm CMOS process in Table 1.5. What is the gain of the stage?

Solution

We first find the transconductance,

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)_1 I_{bias}} = 0.735\ \text{mA/V} \tag{3.11}$$

Also,

$$r_{ds1} = r_{ds2} = \frac{L}{\lambda L I_D} = \frac{0.2\ \mu m}{(0.08\ \mu m/V)(100\ \mu A)} = 25\ k\Omega \tag{3.12}$$

1. Whenever a variable is designated G_i, it is assumed that the variable is an admittance and that $G_i = 1/R_i$, where R_i is the resistance of the same component.

The equation for the body-effect parameter, from Chapter 1, is

$$g_{s1} = \frac{\gamma g_m}{2\sqrt{V_{SB} + |2\phi_F|}} \tag{3.13}$$

To calculate this parameter, we need to know the source-bulk voltage, V_{SB}. Unfortunately, this voltage is dependent on the application and cannot be known accurately beforehand. Here we will assume that $V_{SB} \approx 0.5$ V to obtain an estimate. We therefore have

$$g_{s1} = \frac{0.3\ \text{V}^{-1/2} \cdot g_m}{2\sqrt{0.5\ \text{V} + 0.8\ \text{V}}} \cong 0.13 g_m = 0.1\ \text{mA/V} \tag{3.14}$$

Using (3.10), we have

$$A_V = \frac{0.735\ \text{mA/V}}{0.735\ \text{mA/V} + 0.1\ \text{mA/V} + 0.04\ \text{mA/V} + 0.04\ \text{mA/V}} = 0.8\ \text{V/V} = -1.9\ \text{dB} \tag{3.15}$$

Note that, as mentioned above, the body-effect parameter, g_{s1} is larger than the other parasitic conductances g_{ds1} and g_{ds2} and, hence, plays a dominant role in limiting the gain. If the body effect were not present (for example, if the source and body could be shorted together) the gain would be increased to around –0.9 dB.

3.4 COMMON-GATE AMPLIFIER

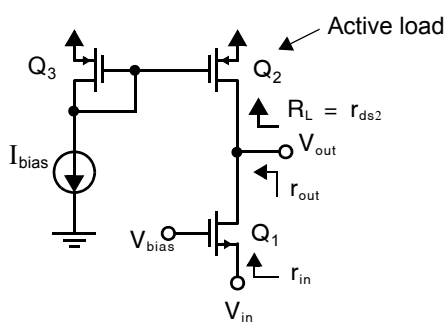


Fig. 3.9 A common-gate amplifier with a current-mirror active load.

A common-gate amplifier with an active load is shown in Fig. 3.9. This stage is commonly used as a gain stage when a relatively small input impedance is desired. For example, it might be designed to have an input impedance of 50 Ω to terminate a 50-Ω transmission line. Another common application for a common-gate amplifier is as the first stage of an amplifier where the input signal is a current; in such cases a small input impedance is desired in order to ensure all of the current signal is drawn into the amplifier, and none is "lost" in the signal source impedance. Aside from its low input impedance, the common-gate amplifier is similar to a common-source amplifier; in both cases the input is applied across v_{gs}, except with opposite polarities, and the output is taken at the drain. Hence, in both cases the small signal gain magnitude approximately equals the product of g_m and the total impedance at the drain.

If we use straightforward small-signal analysis, when the impedance seen at V_{out} (in this case, the output impedance of the current mirror formed by Q_2) is much less than r_{ds1}, the input impedance, r_{out}, is found to be $1/g_{m1}$ at low frequencies. However, in integrated applications, the impedance seen at V_{out} is often on the same order of magnitude or even much greater than r_{ds1}. In this case, the input impedance at low frequencies can be considerably larger than $1/g_{m1}$. To see this result, consider the small-signal model shown in Fig. 3.10. In this model, the voltage-dependent current source that models the body effect has been included. Notice that $v_{gs1} = -v_{s1}$ and therefore the two current sources can be combined into a single current source, as shown in Fig. 3.11. This simplification is always possible for a transistor that has a grounded gate in a small-signal model, and considerably simplifies taking the body effect

into account. Specifically, one can simply ignore the body effect for transistors with grounded gates, and then, after the analysis is complete, simply replace the constants g_{mi} with $g_{mi} + g_{si}$. However, for this example, we include the body-effect parameter throughout the analysis.

At node v_{out}, we have

$$v_{out}(G_L + g_{ds1}) - v_{s1}g_{ds1} - (g_{m1} + g_{s1})v_{s1} = 0 \tag{3.16}$$

Rearranging slightly, we have

$$\frac{v_{out}}{v_{s1}} = \frac{g_{m1} + g_{s1} + g_{ds1}}{G_L + g_{ds1}} = (g_{m1} + g_{s1} + g_{ds1})(R_L \parallel r_{ds1}) \cong g_{m1}(R_L \parallel r_{ds1}) \tag{3.17}$$

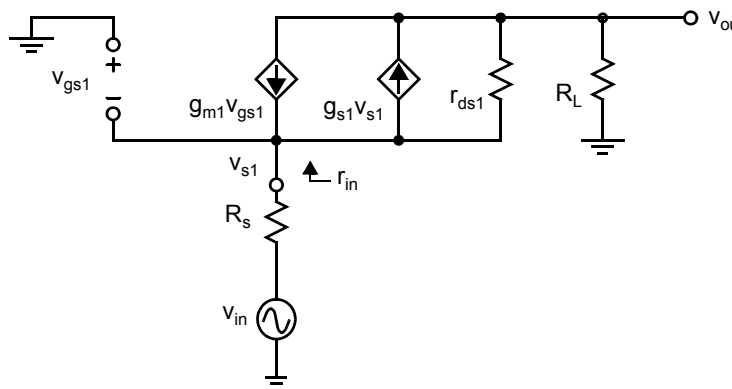


Fig. 3.10 The small-signal model of the common-gate amplifier at low frequencies.

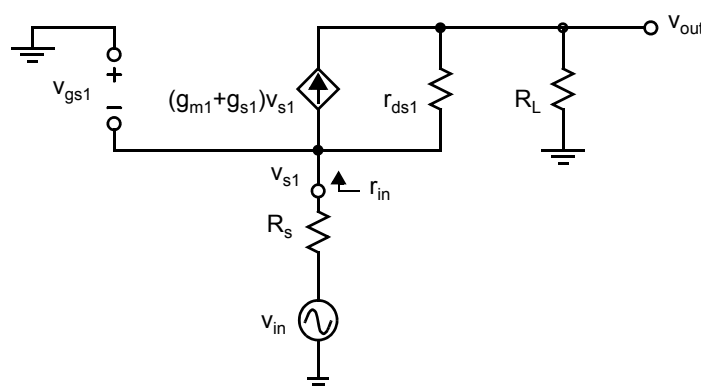


Fig. 3.11 A simplified small-signal model of the common-gate amplifier.

The current going into the source of Q_1 is given by

$$i_s = v_{s1}(g_{m1} + g_{s1} + g_{ds1}) - v_{out}g_{ds1} \tag{3.18}$$

Combining (3.17) and (3.18) to find the input admittance, $g_{in} = 1/r_{in}$, we have

$$g_{in} \equiv \frac{i_s}{v_{s1}} \equiv \frac{g_{m1} + g_{s1} + g_{ds1}}{1 + \dfrac{g_{ds1}}{G_L}} \cong \frac{g_{m1}}{1 + \dfrac{g_{ds1}}{G_L}} \tag{3.19}$$

Alternatively, we have

$$r_{in} = \frac{1}{g_{in}} = \left(\frac{1}{g_{m1}} \parallel \frac{1}{g_{s1}} \parallel r_{ds1}\right)\left(1 + \frac{R_L}{r_{ds1}}\right) \cong \frac{1}{g_{m1}}\left(1 + \frac{R_L}{r_{ds1}}\right) \tag{3.20}$$

Key Point: *The common-gate amplifier provides a voltage gain comparable to that of the common-source amplifier, but with a relatively low input resistance on the order of* $1/g_m$. *However, the input resistance can be larger when the amplifier has a large small-signal load resistance.*

With the p-channel active load shown in Fig. 3.9, $R_L = r_{ds2}$. Since, in this case, R_L is approximately the same magnitude as r_{ds1}, the input impedance, r_{in}, is about $2/g_{m1}$ for low frequencies—twice as large as the expected value of $1/g_{m1}$. This increased input impedance must be taken into account in applications such as transmission-line terminations. In some examples, the current-mirror output impedance realized by Q_2 is much larger than r_{ds1} (i.e., $R_L >> r_{ds1}$), and so *the input impedance for this common-gate amplifier is much lar ger than* $1/g_{m1}$. This increased input impedance often occurs in integrated circuits and is not commonly known.

The attenuation from the input source to the transistor source can be considerable for a common-gate amplifier when R_s is large. This attenuation is given by

$$\frac{v_{s1}}{v_{in}} = \frac{r_{in}}{R_s + r_{in}} \tag{3.21}$$

Using (3.20) to replace r_{in}, we have

$$\frac{v_{s1}}{v_{in}} = \frac{\left(\dfrac{1}{g_{m1}} \parallel \dfrac{1}{g_{s1}} \parallel r_{ds1}\right)\left(1 + \dfrac{R_L}{r_{ds1}}\right)}{R_s + \left(\dfrac{1}{g_{m1}} \parallel \dfrac{1}{g_{s1}} \parallel r_{ds1}\right)\left(1 + \dfrac{R_L}{r_{ds1}}\right)} = \frac{1}{1 + R_s\left(\dfrac{g_{m1} + g_{s1} + g_{ds1}}{1 + R_L/r_{ds1}}\right)} \tag{3.22}$$

Using (3.17) and (3.22), we find that the overall dc gain is given by

$$A_V = \frac{v_{out}}{v_{in}} = \frac{(g_{m1} + g_{s1} + g_{ds1})(R_L \parallel r_{ds1})}{1 + R_s\left(\dfrac{g_{m1} + g_{s1} + g_{ds1}}{1 + R_L/r_{ds1}}\right)} \cong \frac{g_{m1}(R_L \parallel r_{ds1})}{1 + R_s\left(\dfrac{g_{m1} + g_{s1} + g_{ds1}}{1 + R_L/r_{ds1}}\right)} \tag{3.23}$$

EXAMPLE 3.5

Design the common-gate amplifier of Fig. 3.9 to have an input impedance of approximately 50 Ω using the 0.18-μm CMOS devices in Table 1.5 with $I_{bias} = 100\ \mu A$, $(W/L)_3 = 2\ \mu m/0.2\ \mu m$.

Solution

Since $(\lambda L)_1 = (\lambda L)_2$, if we take $L_1 = L_2 = 0.2\ \mu m$ we ensure $r_{ds2} = R_L = r_{ds1}$ and using (3.20),

$$r_{in} \cong 2/g_{m1} \tag{3.24}$$

Hence, to ensure $r_{in} = 50\ \Omega$,

$$g_{m1} \cong 2/(50\ \Omega) = 40\ mA/V \tag{3.25}$$

If we take $V_{eff1} = 200\ mV$,

$$I_{D2} = I_{D1} = g_{m1}V_{eff1}/2 = (40\ mA/V)(200\ mV)/2 = 4\ mA \tag{3.26}$$

This requires the current mirror to provide a gain of 4 mA/100 μA = 40.

$$(W/L)_2 = 40(W/L)_3 = 80\ \mu m/0.2\ \mu m$$

Finally, using the values in (3.25) and (3.26),

$$(W/L)_1 = \frac{g_{m1}^2}{2\mu_n C_{ox} I_{D1}} = 148\ \mu m/0.2\ \mu m$$

In reality, the input impedance will be somewhat lower than 50 Ω due to the body effect and finite drain-source resistance, r_{ds1}, not included in (3.24).

3.5 SOURCE-DEGENERATED CURRENT MIRRORS

We saw in Section 3.1 that a current mirror can be realized using only two transistors, where the output impedance of this current source was seen to be r_{ds2}. To increase this output impedance, a source-degenerated current mirror can be used, as shown in Fig. 3.12. The small-signal model for this current mirror is shown in Fig. 3.13. Since no small-signal current flows into the gate, the gate voltage is 0 V in Fig. 3.13.

Note that the current i_x sourced by the applied voltage source is equal to the current through the degeneration resistor, R_s. Therefore, we have

$$v_s = i_x R_s \tag{3.27}$$

Also, note that

$$v_{gs} = -v_s \tag{3.28}$$

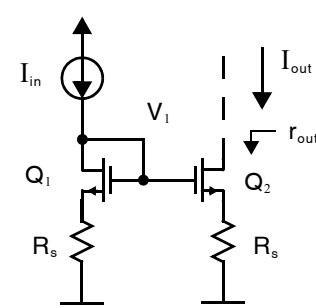


Fig. 3.12 A current mirror with source degeneration.

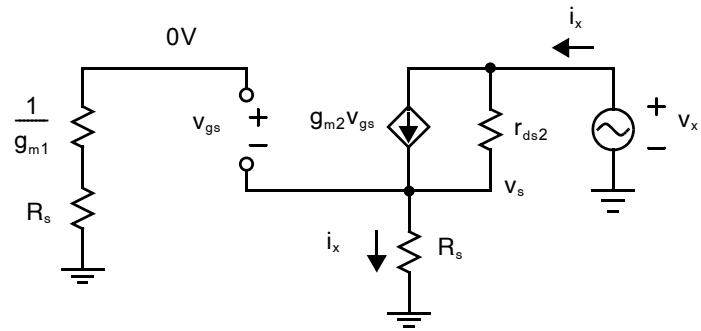


Fig. 3.13 The small-signal model for the source-degenerated current source.

Setting i_x equal to the total current through $g_{m2}v_{gs}$ and r_{ds2} gives

$$i_x = g_{m2}v_{gs} + \frac{v_x - v_s}{r_{ds2}} \tag{3.29}$$

Substituting (3.27) and (3.28) into (3.29) gives

$$i_x = -i_x g_{m2} R_s + \frac{v_x - i_x R_s}{r_{ds2}} \tag{3.30}$$

Rearranging, we find the output impedance to be given by

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{ds2})] \cong r_{ds2}(1 + R_s g_{m2}) \tag{3.31}$$

where g_{ds2} is equal to $1/r_{ds2}$, which is much less than g_{m2}. Thus, the output impedance has been increased by a factor approximately equal to $(1 + R_s g_{m2})$.

Key Point: *When a small-signal resistance R_s is introduced at the source of both transistors in a simple current mirror, the output resistance is increased by a factor approximately equal to* $(1 + R_s g_m)$.

This formula can often be applied to moderately complicated circuits to quickly estimate the impedances looking into a node. Such an example follows, in the derivation of the output impedance of cascode current mirrors. It should be noted that the above derivation ignores the body effect of the transistor, even though the source of the transistor is not connected to a small-signal ground. As discussed earlier, in Section 3.4, since the gate is at a small-signal ground, the body effect can be taken into account by simply replacing g_{m2} in Eq. (3.31) with $g_{m2} + g_{s2}$. This substitution results in

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{s2} + g_{ds2})] \cong r_{ds2}[1 + R_s(g_{m2} + g_{s2})] \tag{3.32}$$

where g_{s2} is the body-effect constant. This result is only slightly different since g_s is roughly one-fifth of g_m.

EXAMPLE 3.6

Consider the current mirror shown in Fig. 3.12, where $I_{in} = 100\ \mu A$, each transistor has $W/L = 10\ \mu m/0.2\ \mu m$, and $R_s = 1\ k\Omega$. Using the 0.18-μm CMOS devices in Table 1.5, find r_{out} for the current mirror. Assume the body effect can be approximated by $g_s = 0.15 g_m$.

Solution

Nominally, $I_{out} = I_{in}$, and thus we find the small-signal parameters for this current mirror to be

$$g_{m2} = \sqrt{2\mu_n C_{ox}(W/L)I_{out}} = 1.64 \text{ mA/V} \tag{3.33}$$

Also we have

$$r_{ds2} = \frac{0.2 \ \mu\text{m}}{0.08 \ \mu\text{m/V} \ 100 \ \mu\text{A}} = 25 \text{ k}\Omega \tag{3.34}$$

Now, making use of (3.32), the output impedance is given by

$$r_{out} = 25 \text{ k}\Omega\left[1 + 1 \text{ k}\Omega\left(1.64 \text{ mA/V} + 0.15 \cdot 1.64 \text{ mA/V} + \frac{1}{25 \text{ k}\Omega}\right)\right] = 73.15 \text{ k}\Omega \tag{3.35}$$

Note that this result is nearly three times the output impedance for a simple current mirror which would be simply $r_{ds2} = 25$ kΩ. Also note that the voltage drop across R_s equals 100 μA × 1 kΩ = 0.1 V due to the dc bias current through it.

3.6 CASCODE CURRENT MIRRORS

A cascode current mirror is shown in Fig. 3.14. First, note that the output impedance looking into the drain of Q_2 is simply r_{ds2}, which is seen using an analysis very similar to that which was used for the simple current mirror. Thus, the output impedance can be immediately derived by considering Q_4 as a current source with a source-degeneration resistor of value r_{ds2}. Making use of (3.32), and noting that Q_4 is now the cascode transistor rather than Q_2, we have

$$r_{out} = r_{ds4}[1 + R_s(g_{m4} + g_{s4} + g_{ds4})] \tag{3.36}$$

where now $R_s = r_{ds2}$. Therefore, the output impedance is given by

$$\begin{aligned} r_{out} &= r_{ds4}[1 + r_{ds2}(g_{m4} + g_{s4} + g_{ds4})] \\ &\cong r_{ds4}[1 + r_{ds2}(g_{m4} + g_{s4})] \\ &\cong r_{ds4}(r_{ds2}g_{m4}) \end{aligned} \tag{3.37}$$

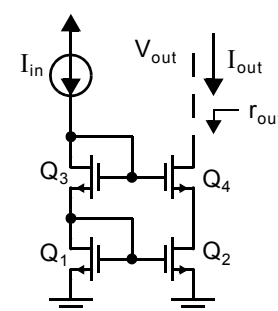


Fig. 3.14 A cascode current mirror.

Thus, the output impedance has been increased by a factor of $g_{m4}r_{ds4}$, which is an upper limit on the gain of a single-transistor MOS gain-stage, and might be a value between 10 and 100, depending on the transistor sizes and currents and the technology being used. This significant increase in output impedance can be instrumental in realizing single-stage amplifiers with large low-frequency gains.

The disadvantage in using a cascode current mirror is that it reduces the maximum output voltage swings possible before transistors enter the triode region. To understand this reduction, recall that for an n-channel transistor to be in the active region (also called the saturation or pinch-off region) its drain-source voltage must be greater than

Key Point: *The addition of a cascode device to a CMOS current mirror increases its output resistance by approximately the gain of the cascode device, $g_m r_{ds}$. The voltage swing available at the current mirror output is reduced because some voltage drop must be maintained across the cascode device to keep it in active mode.*

$$V_{eff} \equiv V_{GS} - V_{tn} \tag{3.38}$$

which was shown in Chapter 1 to be given by

$$V_{eff} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{3.39}$$

If we assume all transistors have the same sizes and currents, then they also all have the same V_{eff} and, therefore, the same gate-source voltages, $V_{GSi} = V_{eff} + V_{tn}$. Also, from Fig. 3.14, we see that

$$V_{G3} = V_{GS1} + V_{GS3} = 2V_{eff} + 2V_{tn} \tag{3.40}$$

and

$$V_{DS2} = V_{G3} - V_{GS4} = V_{G3} - (V_{eff} + V_{tn}) = V_{eff} + V_{tn} \tag{3.41}$$

Thus, the drain-source voltage of Q_2 is larger than the minimum needed to place it at the edge of the active region. Specifically, the drain-source voltage of Q_2 is V_{tn} greater than what is required. Since the smallest output voltage, V_{D4}, can be without Q_4 entering the triode region is given by $V_{DS2} + V_{eff}$, the minimum allowed voltage for V_{out} is given by

$$V_{out} > V_{DS2} + V_{eff} = 2V_{eff1} + V_{tn} \tag{3.42}$$

which, again, is V_{tn} greater than the minimum value of $2V_{eff}$. This loss of signal swing is a serious disadvantage when modern technologies are used that might have a maximum allowed power-supply voltage as small as 1 V. In Section 6.3, we will see how the cascode current mirror can be modified to maintain large output impedances and yet still allow for near minimum voltages at the output of the mirror.

EXAMPLE 3.7

Consider the cascode current mirror shown in Fig. 3.14, where $I_{in} = 100\ \mu A$ and each transistor has $W/L = 10\ \mu m/0.4\ \mu m$. Given the 0.35-μm CMOS device parameters in Table 1.5, find r_{out} for the current mirror (approximating the body effect by $0.2g_m$). Also find the minimum output voltage at V_{out} such that the output transistors remain in the active region.

Solution

Nominally, $I_{out} = I_{in}$, and thus we can find the small-signal parameters for this current mirror to be

$$g_{m4} = \sqrt{2\mu_n C_{ox}(W/L)I_{out}} = 0.97\ mA/V \tag{3.43}$$

We also have

$$r_{ds2} = r_{ds4} = \frac{0.4\ \mu m}{0.16\ \mu m/V\ 100\ \mu A} = 25\ k\Omega \tag{3.44}$$

Now, making use of (3.37), the output impedance is given by

$$r_{out} = 25\ k\Omega[25\ k\Omega(0.97\ mA/V + 0.2 \times 0.97\ mA/V + 1/25\ k\Omega)] = 753\ k\Omega \tag{3.45}$$

To find the minimum output voltage, we first need to determine V_{eff}:

$$V_{eff} = \sqrt{\frac{2I_{out}}{\mu_n C_{ox}(W/L)}} = 0.205\ V \tag{3.46}$$

Thus, the minimum output voltage is determined to be $2V_{eff} + V_{tn} = 0.98$ V. Hence, compared with the simple current mirror with the same current and transistors in Example 3.1, the output resistance is increased by a factor of 30, but the voltage swing available at the output is decreased by 0.76 V.

3.7 CASCODE GAIN STAGE

In modern IC design, a commonly used configuration for a single-stage amplifier is a cascode configuration. This configuration consists of a common-source-connected transistor feeding into a common-gate-connected transistor. Two examples of cascode amplifiers are shown in Fig. 3.15. The configuration in Fig. 3.15(*a*) has both an n-channel common-source transistor, Q_1, and an n-channel common-gate cascode transistor, Q_2. This configuration is sometimes called a telescopic-cascode amplifier. The configuration shown in Fig. 3.15(*b*) has an n-channel input (or drive) transistor, but a p-channel transistor is used for the cascode (or common-gate) transistor. This configuration is usually called a folded-cascode stage. When incorporated into an operational amplifier, the folded cascode can provide greater output swing than the telescopic cascode. However, the folded cascode usually consumes more power since the drain bias currents for Q_1 and Q_2 are drawn in parallel from the supply voltage, whereas in the telescopic cascode the same dc drain current is shared by both transistors.

There are two major reasons for the popularity of cascode stages. The first is that they can have quite large gain for a single stage due to the large impedances at the output. To enable this high gain, the current sources connected to the output nodes are realized using high-quality cascode current mirrors. The second major reason for the use of cascode stages is that they limit the voltage across the input drive transistor. This minimizes any short-channel effects, which becomes more important with modern technologies having very short channel-length transistors. It can also permit the circuit to handle higher output voltages without risking damage to the common-source transistor.

Key Point: *The cascode gain stage uses a common-gate transistor Q_2 to reduce the V_{DS} variations on a common-source transistor Q_1. The result is high output resistance providing potentially high gain and reduced short-channel effects. However, the available output voltage swing is reduced.*

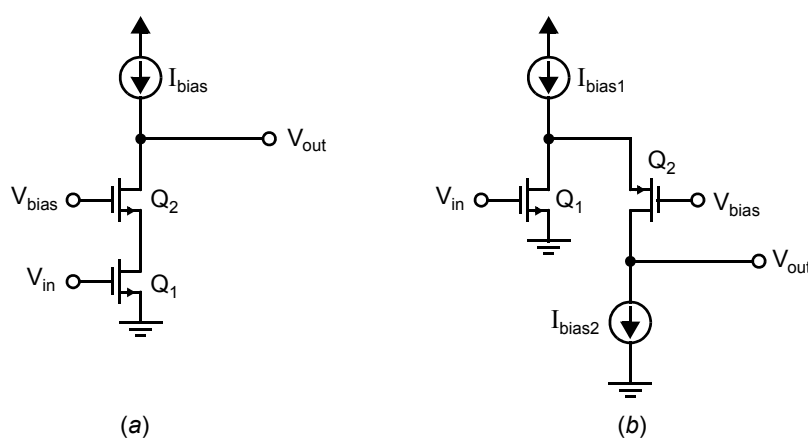


Fig. 3.15 (*a*) A telescopic-cascode amplifier and (*b*) a folded-cascode amplifier.

The main drawback of cascode amplifiers is that the output voltage is restricted to a narrower range than the common-source amplifier in order to keep all devices in the active region. This is because some voltage must be kept between drain and source of the extra cascode transistor, Q_2.

EXAMPLE 3.8

For the telescopic cascode gain stage pictured in Fig. 3.15*(a)*, what is the minimum voltage that can appear at v_{out} while keeping both Q_1 and Q_2 in the active region? How does this compare with the common-source amplifier in, for example, Fig. 3.4? Assume that all transistors are biased with $V_{eff} = 250\text{mV}$.

Solution

In order to keep Q_1 in active mode, a voltage of at least V_{eff} is required at the drain of Q_1. Hence,

$$V_{bias} \geq V_{eff,1} + V_{GS,2} = 2V_{eff} + V_{tn} \tag{3.47}$$

The absolute minimum voltage possible at V_{out} is achieved by taking the minimum possible value for V_{bias} allowed by (3.47). Since V_{out} may be at most one threshold voltage below V_{bias}, it is required that

$$V_{out} \geq 2V_{eff} = 500\text{mV} \tag{3.48}$$

In practice, a significant margin is provided beyond this value since the exact values of V_{tn} and V_{eff} are not known.

In the common-source amplifier of Fig. 3.4, there is only one transistor between V_{out} and ground (Q_1). Hence, the output can drop to within one V_{eff} of ground, or just 250mV. The additional 250mV of swing provided by the common-source amplifier compared with a cascode amplifier can be very significant when operating with low supply voltages. This is a major reason why common-source amplifiers are often used as the output stage of operational amplifiers.

The small-signal analysis of the telescopic cascode gain stage of Fig. 3.15(*a*) will next be described. The same analysis, with only minor modifications, also applies for the folded-cascode stage of Fig. 3.15(*b*). A small-signal schematic for the analysis is shown in Fig. 3.16, where the transistor symbols remain as shorthand for their complete small-signal models.

It is useful to know the small-signal resistances at several points in the circuit, which are defined in Fig. 3.16. From the section on cascode current-mirrors, we know that the impedance looking into the drain of cascode

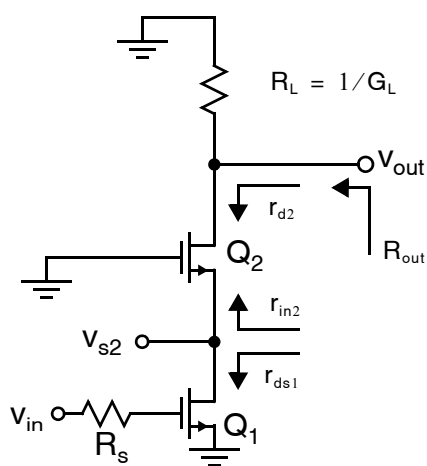


Fig. 3.16 A small-signal equivalent circuit of the telescopic cascode amplifier where MOS transistor symbols are used as shorthand for their small-signal models.

transistor Q_2 is approximately given by

$$r_{d2} \cong g_{m2} r_{ds1} r_{ds2} \tag{3.49}$$

The total output resistance will be

$$R_{out} = r_{d2} \| R_L \tag{3.50}$$

where R_L is the output impedance of the bias current source, I_{bias}. We found earlier that the low-frequency admittance looking into the source of the common-gate or cascode transistor, Q_2, is[2]

$$g_{in2} = 1/r_{in2} = \frac{g_{m2} + g_{s2} + g_{ds2}}{1 + \dfrac{R_L}{r_{ds2}}} \cong \frac{g_{m2}}{1 + \dfrac{R_L}{r_{ds2}}} \tag{3.51}$$

The gain from the input to the source of Q_2 is simply that of a common-source amplifier with a load resistance of r_{in2} and is therefore given by

$$\frac{v_{s2}}{v_{in}} = -g_{m1}(r_{ds1} \| r_{in2}) = -\frac{g_{m1}}{g_{ds1} + g_{in2}} \tag{3.52}$$

The gain from the source of Q_2 to the output is simply that derived earlier for the common-gate stage.

$$\begin{aligned} \frac{v_{out}}{v_{s2}} &= \frac{g_{m2} + g_{s2} + g_{ds2}}{g_{ds2} + G_L} \\ &\cong g_{m2}(r_{ds2} \| R_L) = \frac{g_{m2}}{(g_{ds2} + G_L)} \end{aligned} \tag{3.53}$$

The overall gain is the product of (3.52) and (3.53).

$$A_V = \frac{v_{s2}}{v_{in}} \frac{v_{out}}{v_{s2}} \cong -g_{m1} g_{m2} (r_{ds1} \| r_{in2})(r_{ds2} \| R_L) \tag{3.54}$$

The reader should be cautioned that (3.54) is only approximate primarily due to the difficulty of accurately determining the output resistance r_{ds} for the different transistors. For example, one problem in estimating r_{ds} is that it is voltage-dependent. Therefore, prudent designers should never create a design where successful operation is dependent on knowing the gain accurately rather than just knowing it will be greater than some minimum value.

EXAMPLE 3.9

Find approximate expressions for the gain and output resistance of the cascode stage in Fig. 3.15(*a*) assuming I_{bias} is a simple current source with an output impedance on the order of

$$R_L \approx r_{ds\text{-}p} \tag{3.55}$$

Compute approximate numerical results assuming all transistors have g_m on the order of 0.5mA/V and r_{ds} on the order of 100kΩ.

2. Compared to the common-gate analysis, the indices are changed to reflect the fact that Q_2 is the common-gate transistor rather than Q_1.

Solution

We shall drop the indices for all small-signal values under the assumption that the transistors are somewhat matched and to simplify matters since we are only deriving an approximate answer.

The resistance looking into the source of Q_2 is obtained by substituting (3.55) into (3.51).

$$g_{in2} \approx \frac{1}{2} g_m \Rightarrow r_{in2} \approx \frac{2}{g_m} \tag{3.56}$$

Assuming r_{ds1} is much larger than r_{in2}, the gain from the input to the source of Q_2 simplifies to

$$\frac{v_{s2}}{v_{in}} \approx -\frac{g_m}{g_m/2} = -2 \tag{3.57}$$

For instance, compared with a common-source amplifer assuming the same numerical values for the transistor small-signal parameters, (v_{s2}/v_{in}) has decreased from –25 to –2. The gain of the common-gate stage also decreases. Substituting (3.55) into (3.53) gives

$$\frac{v_{out}}{v_{s2}} \approx \frac{1}{2} g_m r_{ds} \tag{3.58}$$

Hence, the overall gain is,

$$A_V = \frac{v_{s2}}{v_{in}} \cdot \frac{v_{out}}{v_{s2}} \approx -g_m r_{ds} \tag{3.59}$$

Since $R_L \ll r_{d2}$, the output resistance is now dominated by (3.55).

$$R_{out} \approx r_{ds} \tag{3.60}$$

The corresponding numerical results are $A_V \approx -50$ or 34 dB and $R_{out} \approx 100\text{k}\Omega$, The gain in this case is, in fact, only a factor of 2 larger than would be obtained with a common-source amplifier (i.e. excluding Q_2). Notice that almost all of the gain appears across the common-gate transistor, Q_2.

The cascode amplifier is most often used in analog integrated circuits to provide a large low-frequency gain from a single stage. Considering equation (3.50), this may only be done if R_L is large. Since R_L is the output resistance of the current source, I_{bias} should be a high-quality current source whose output resistance is on the order of $R_L \approx g_{m\text{-}p} r_{ds\text{-}p}^2$.

We shall now obtain expressions for the gain and output resistance in this case using the results of our previous analysis and dropping all indices to obtain an approximate result. Substituting $R_L \approx g_m r_{ds}^2$ into (3.51) reveals that the conductance looking into the source of Q_2 is

$$\begin{aligned} g_{in2} = 1/r_{in2} &\approx \frac{g_m}{1 + g_{ds} g_m r_{ds}^2} \cong g_{ds} \\ &\Rightarrow r_{in2} \approx r_{ds} \end{aligned} \tag{3.61}$$

Therefore, the gain from the input to v_{s2} in (3.52) may be simplified to,

$$\frac{v_{s2}}{v_{in}} \approx -\frac{1}{2} g_m r_{ds} \tag{3.62}$$

and since $R_L \gg r_{ds2}$, the gain from v_{s2} to the output in (3.53) becomes simply

$$\frac{v_{out}}{v_{s2}} \approx g_m r_{ds} \tag{3.63}$$

The approximate overall gain is the product of (3.62) and (3.63),

$$A_V \approx -\frac{1}{2} g_m^2 r_{ds}^2 = -\frac{1}{2}\left(\frac{g_m}{g_{ds}}\right)^2 \tag{3.64}$$

The output resistance is

$$R_{out} = R_L \| r_{d2} \approx \frac{1}{2} g_m r_{ds}^2 \tag{3.65}$$

EXAMPLE 3.10

Compare the gain and output resistance of the cascode amplifier with a high-quality current source to the values obtained in Example 3.9 where a simple current source was assumed. As in Example 3.9, assume all transistors have g_m on the order of 0.5mA/V and r_{ds} on the order of $100\text{k}\Omega$.

Solution

Substituting the numerical values into equation (3.65) yields $A_V \approx -1250$, a very large gain from a single amplifier stage. The output resistance is huge, $R_{out} \approx 2.5\text{M}\Omega$. These are both over an order of magnitude larger than the values $A_V \approx -50$ and $R_{out} \approx 100\text{k}\Omega$ obtained in Example 3.9; the values have been effectively magnified by the gain of the common-gate transistor Q_2.

3.8 MOS DIFFERENTIAL PAIR AND GAIN STAGE

Most integrated amplifiers have a differential input. To realize this differential input, almost all amplifiers use what is commonly called a differential transistor pair. A differential pair together with a biasing current source is shown in Fig. 3.17. The transistors Q_1 and Q_2 are sized identically and are biased with the same dc gate voltage. A low-frequency, small-signal model of the differential pair is shown in Fig. 3.18. This small-signal equivalent circuit is based on the T model for a MOS transistor that was described in Chapter 1. To simplify the analysis, we ignore the output impedance of the transistors temporarily. Defining the differential input voltage as $v_{in} \equiv v^+ - v^-$, we have

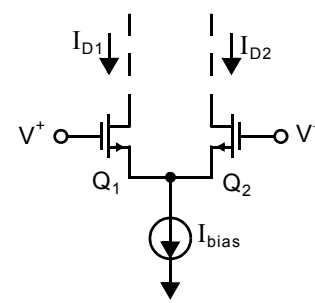


Fig. 3.17 A MOS differential pair.

$$i_{d1} = i_{s1} = \frac{v_{in}}{r_{s1} + r_{s2}} = \frac{v_{in}}{1/g_{m1} + 1/g_{m2}} \tag{3.66}$$

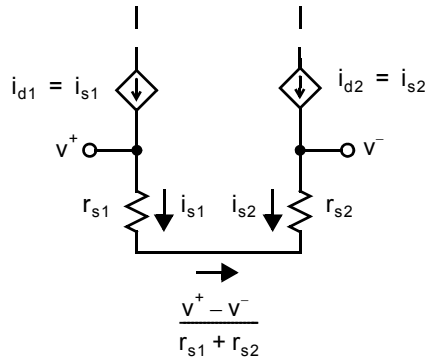


Fig. 3.18 The small-signal model of a MOS differential pair.

Since both Q_1 and Q_2 have the same bias currents, $g_{m1} = g_{m2}$. Therefore, we find

$$i_{d1} = \frac{g_{m1}}{2} v_{in} \tag{3.67}$$

Also, since $i_{d2} = i_{s2} = -i_{d1}$, we find that

$$i_{d2} = -\frac{g_{m1}}{2} v_{in} \tag{3.68}$$

Finally, defining a differential output current, $i_{out} \equiv i_{d1} - i_{d2}$, the following relationship is obtained:

$$i_{out} = g_{m1} v_{in} \tag{3.69}$$

Key Point: *The differential pair with active load is classically the first stage of a two-stage operational amplifier. It accepts a differential input and produces a single-ended output with a gain comparable to that of a common-source stage.*

If two resistive loads R_L are connected between the drains of Q_1 and Q_2 and a positive supply, the result is a differential output voltage between the two drain nodes, $v_{out} = (g_{m1}R_L)v_{in}$ and the stage has a differential small-signal gain of $g_{m1}R_L$.

If a differential pair has a current mirror as an active load, a complete differential-input, single-ended-output gain stage can be realized, as shown in Fig. 3.19. This circuit is the typical first gain stage of a classical two-stage integrated opamp; in this case, the input differential pair is realized using n-channel transistors and the active current-mirror load is realized using p-channel transistors. From the small-signal analysis of the differential pair, we have

$$i_{d1} = i_{s1} = \frac{g_{m1}}{2} v_{in} \tag{3.70}$$

Also, ignoring transistor output impedances, we have

$$i_{d4} = i_{d3} = -i_{s1} \tag{3.71}$$

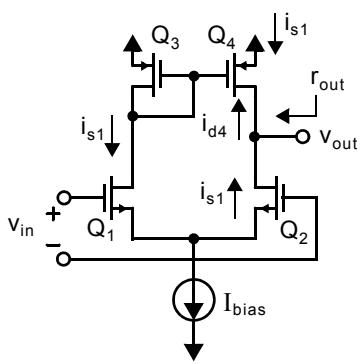


Fig. 3.19 A differential-input, single-ended-output MOS gain stage.

Note that a positive small-signal current is defined as the current going into the drain of a transistor. Using (3.71) and the fact that $i_{d2} = -i_{s1}$, we have

$$v_{out} = (-i_{d2} - i_{d4})r_{out} = 2i_{s1}r_{out} = g_{m1}r_{out}v_{in} \tag{3.72}$$

The evaluation of the output resistance, r_{out}, is determined by using the small-signal equivalent circuit and applying a voltage to the output node, as seen in Fig. 3.20. Note that the T model was used for both Q_1 and Q_2, whereas Q_3 was replaced by an equivalent resistance (since it is diode-connected), and the hybrid-π model was used for Q_4. As usual, r_{out} is defined as the ratio v_x/i_x, where i_x is given by the sum $i_x = i_{x1} + i_{x2} + i_{x3} + i_{x4}$. Clearly,

$$i_{x1} = \frac{v_x}{r_{ds4}} \tag{3.73}$$

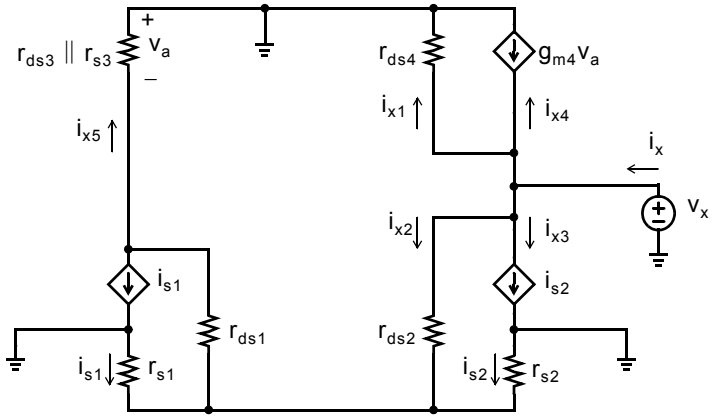


Fig. 3.20 The small-signal model for the calculation of the output impedance of the differential-input, single-ended-output MOS gain stage.

implying that the resistance seen in the path taken by i_{x1} is equal to r_{ds4}. Now, assuming that the effect of r_{ds1} can be ignored (since it is much larger than r_{s1}), we see that the current i_{x2} is given by

$$i_{x2} \cong \frac{v_x}{r_{ds2} + (r_{s1} \parallel r_{s2})} \cong \frac{v_x}{r_{ds2}} \tag{3.74}$$

where the second approximation is valid, since r_{ds2} is typically much greater than $r_{s1} \parallel r_{s2}$. This i_{x2} current splits equally between i_{s1} and i_{s2} (assuming $r_{s1} = r_{s2}$ and once again ignoring r_{ds1}), resulting in

$$i_{s1} = i_{s2} = \frac{-v_x}{2r_{ds2}} \tag{3.75}$$

However, since the current mirror realized by Q_3 and Q_4 results in $i_{x4} = i_{x5}$ (assuming $g_{m4} = 1/r_{s4} = 1/r_{s3}$ and r_{ds3} is much larger than r_{s3}), the current i_{x4} is given by

$$i_{x4} = -i_{s1} = -i_{s2} = -i_{x3} \tag{3.76}$$

In other words, when the current splits equally between r_{s1} and r_{s2}, the current mirror of Q_3 and Q_4 causes the two currents i_{x3} and i_{x4} to cancel each other. Finally, the output resistance, r_{out}, is given by

$$r_{out} = \frac{v_x}{i_{x1} + i_{x2} + i_{x3} + i_{x4}} = \frac{v_x}{(v_x/r_{ds4}) + (v_x/r_{ds2})} \tag{3.77}$$

which results in the simple relationship

$$r_{out} = r_{ds2} \parallel r_{ds4} \tag{3.78}$$

Therefore, at low frequencies the gain, A_v, is given by

$$A_v = g_{m1}(r_{ds2} \parallel r_{ds4}) \tag{3.79}$$

EXAMPLE 3.11

Consider the gain stage shown in Fig. 3.19, where $I_{bias} = 200\ \mu A$ and all transistors have $W/L = 20\ \mu m/0.4\ \mu m$. Given the transistor parameters for the 0.18 μm technology in Table 1.5, find the output impedance, r_{out}, and the gain from the differential input to the output, V_{out}.

Solution

To find the bias currents, we assume that I_{bias} splits evenly between the two sides of the differential circuit, resulting in

$$I_{D1} = I_{D2} = I_{D3} = I_{D4} = 100\ \mu A \tag{3.80}$$

Therefore, the transconductance of the input transistors is equal to

$$g_{m1} = g_{m2} = \sqrt{2\mu_n C_{ox}(W/L)(I_{bias}/2)} = 1.64\ \text{mA/V} \tag{3.81}$$

The output impedance of Q_2 and Q_4 is given by

$$r_{ds2} = r_{ds4} = \frac{0.4}{0.08 \times 0.1} = 50\ \text{k}\Omega \tag{3.82}$$

Thus, the gain for this stage is

$$A_v \equiv \frac{v_{out}}{v_{in}} = g_{m1}(r_{ds2} \parallel r_{ds4}) = 41\ \text{V/V} \tag{3.83}$$

3.9 KEY POINTS

- A simple CMOS current mirror has a small-signal input resistance of $1/g_{m1}$ and a small-signal output resistance r_{ds2}. [p. 119]
- The common-source amplifier is a popular gain stage, especially when high input impedance is desired. The use of an active load takes advantage of the nonlinear, large-signal transistor current–voltage relationship to provide large small-signal resistances without large dc voltage drops. [p. 120]
- The source-follower provides a voltage gain close to unity, and often limited by the body effect. It can provide a large current gain and it is unilateral so it is often used as a voltage buffer. [p. 122]
- The common-gate amplifier provides a voltage gain comparable to that of the common-source amplifier, but with a relatively low input resistance on the order of $1/g_m$. However, the input resistance can be larger when the amplifier has a large small-signal load resistance. [p. 126]
- When a small-signal resistance R_s is introduced at the source of both transistors in a simple current mirror, the output resistance is increased by a factor approximately equal to $(1 + R_s g_m)$. [p. 128]
- The addition of a cascode device to a CMOS current mirror increases its output resistance by approximately the gain of the cascode device, $g_m r_{ds}$. The voltage swing available at the current mirror output is reduced because some voltage drop must be maintained across the cascode device to keep it in active mode. [p. 129]
- The cascode gain stage uses a common-gate transistor Q_2 to reduce the V_{DS} variations on a common-source transistor Q_1. The result is high output resistance providing potentially high gain and reduced short-channel effects. However, the available output voltage swing is reduced. [p. 131]
- The differential pair with active load is classically the first stage of a two-stage operational amplifier. It accepts a differential input and produces a single-ended output with a gain comparable to that of a common-source stage. [p. 136]

3.10 REFERENCES

P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer. *Analysis and Design of An alog Integrated Circuits,* 5th ed. John Wiley & Sons, New York, 2009.

A. S. Sedra and K. C. Smith. *Microelectronic Circuits,* 6th ed. Holt, Rinehart & Winston, New York, 1991.

3.11 PROBLEMS

3.11.1 Section 3.1: Simple CMOS Current Mirror

3.1 Consider the current mirror shown in Fig. 3.1, where $I_{in} = 80\ \mu A$, transistor Q_1 has $W/L = 100\ \mu m/1.6\ \mu m$, and transistor Q_2 has $W/L = 25\ \mu m/1.6\ \mu m$. Find the nominal output current as well as the output resistance, r_{out}. Also, find the minimum output voltage such that both transistors remain in the active region. Use the NMOS device parameters for the 0.8-μm CMOS process in Table 1.5.

3.2 For the current mirror in Example 3.1, assuming the two transistors are identical, what drain voltage at Q_2 will ensure that I_{in} is precisely equal to I_{out}? Compare your results with SPICE.

3.3 For the current mirror in Example 3.1, find the resulting change in I_{out} if:

a. the threshold voltage of Q_2 increases by 5 mV while that of Q_1 remains the same

b. the carrier mobility of Q_2 increases by 5% while that of Q_1 remains the same

c. the gate length of Q_2 is increased to 0.42 μm while that of Q_1 remains 0.40 μm

Compare your results with SPICE.

3.4 The current mirror in Fig. 3.1 is implemented using the 45-nm NMOS devices in Table 1.5. If $I_{in} = 100\ \mu A$ and $(W/L)_{1,2} = 2\ \mu m/0.1\ \mu m$, estimate the small-signal output resistance of the current mirror.

3.5 Consider the current mirror in Fig. 3.1 with $L_2 = L_1$ and $W_2 = 3W_1$. What is the relationship between I_{out} and I_{in}?

3.6 Consider the current mirror depicted in Fig. P3.6 where $I_{in} = 100\ \mu A$, $R = 2\ k\Omega$, and each transistor has $W/L = 10\ \mu m/0.4\ \mu m$. Given the 0.35-μm CMOS device parameters in Table 1.5, what drain voltage at Q_2 will ensure that I_{in} is precisely equal to I_{out}?

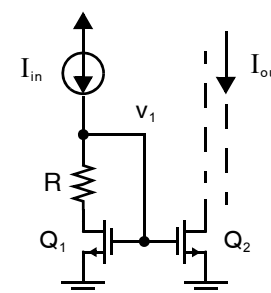


Fig. P3.6

3.7 The three-transistor circuit shown in Fig. P3.7 is to be used as a current source. For this circuit, the 0.18-μm CMOS device parameters in Table 1.5 are used, $(W/L)_1 = 4\ \mu m/0.25\ \mu m$ and $(W/L)_2 = 12\ \mu m/0.25\ \mu m$.

a. Determine the size of transistor Q_3 so that $I_1 = 200\ \mu A$.

b. What is the maximum voltage that can appear at v_o while still keeping Q_3 in saturation?

c. Keeping the device sizes the same, calculate a new value for the resistor so that $I_1 = 250\ \mu A$.

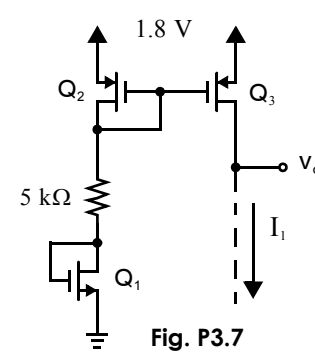


Fig. P3.7

3.11.2 Section 3.2: Common-Source Amplifier

3.8 In Example 3.2, over what range of input voltages will both Q_1 and Q_2 remain in active mode? Assume a 3-V supply voltage.

3.9 Repeat Example 3.2, but doubling all device widths and lengths. How does this effect the result? Compare your results with Spice.

3.10 Design a common-source amplifier with active load using the 0.18-μm CMOS device parameters in Table 1.5 and assuming a supply voltage of 1.8 V. The amplifier should have a power consumption of 1 mW, and a gain of at least 5 V/V. Ensure all transistors have $V_{eff} = 250$ mV and $L = 0.25$ μm.

3.11 Consider the common-source amplifier shown in Fig. P3.11. The transistors are sized $(W/L)_1 = 6\ \mu m/0.3\ \mu m$ and $(W/L)_2 = 4\ \mu m/0.3\ \mu m$. Transistor Q_1 is biased so that $V_{eff1} = 200$ mV. Using the 0.18-μm CMOS device parameters in Table 1.5, and assuming a supply voltage of 1.8 V, estimate:

a. the power consumption of the amplifier

b. the small-signal output resistance of the amplifier

c. the small-signal gain of the amplifier, v_{out}/v_{in}

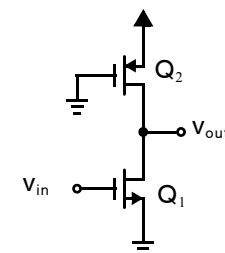


Fig. P3.11

3.12 Consider the common-source amplifier shown in Fig. P3.11. Both devices are sized with $(W/L) = 6\ \mu m/0.3\ \mu m$ and Q_1 is biased so that $V_{eff1} = 200$ mV. Using the 0.18-μm CMOS device parameters in Table 1.5, and assuming a supply voltage of 1.8 V, estimate:

a. the power consumption of the amplifier

b. the small-signal output resistance of the amplifier

c. the small-signal gain of the amplifier, v_{out}/v_{in}

3.13 Draw the schematic of a PMOS common-source amplifier with NMOS current mirror active load. If it is designed using the same bias currents and the same size transistors as the NMOS common-source amplifier in Fig. 3.4, which is likely to have the higher gain? Why?

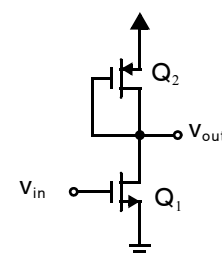


Fig. P3.12

3.11.3 Section 3.3: Source-Follower or Common-Drain Amplifier

3.14 Derive the output resistance of the source follower shown in Fig. 3.6.

3.15 The source follower in Fig. 3.6 is to be designed using devices from the 0.18-μm CMOS process in Table 1.5 to drive a 3 kΩ load resistor connected between v_{out} and ground (not shown in Fig. 3.6). $I_{bias} = 0.2$ mA and all transistors are sized identically with a gate length $L = 0.25$ μm. Choose the gate width W to provide a small-signal gain of at least 0.8 V/V. You may assume that the body of transistor Q_1 is tied to its source so that there is no body effect.

3.11.4 Section 3.4: Common-Gate Amplifier

3.16 Assume that the common-gate amplifier of Fig. 3.9 has a bias current of 0.1 mA and that all transistors have a W/L of 10 μm/0.18 μm. Find the small-signal gain and input resistance of the amplifier. Use the device parameters for the 0.18-μm CMOS process in Table 1.5 and compare your results with Spice.

3.17 The common-gate amplifier of Fig. 3.9 is to be designed to have an input resistance of 200 Ω using the device parameters for the 0.35-μm CMOS process in Table 1.5. All transistors are to have gate lengths of $L = 0.4\ \mu m$ and $V_{eff} = 0.25$ V. Estimate the required values of I_{bias} and $W_{1,2,3}$.

3.11.5 Section 3.5: Source-Degenerated Current Mirrors

3.18 The source-degenerated current mirror in Fig. 3.12 is implemented in the 0.18-μm CMOS process in Table 1.5 to mirror a current of $I_{in} = 0.5$ mA with $(W/L)_{1,2} = 20\ \mu m/0.2\ \mu m$.

a. Estimate the small-signal output resistance and minimum output voltage of the mirror with $R_s = 200\ \Omega$.

b. Compare the results in part (a) to those is obtained with $R_s = 0$, a simple NMOS current mirror.

3.19 Derive the output resistance of the current mirror shown in Fig. P3.19 where a diode-connected transistor has been included in series with the source of the output transistor. Ignore the body effect.

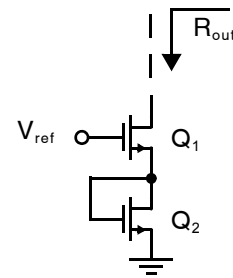


Fig. P3.19

3.11.6 Section 3.6: Cascode Current Mirrors

3.20 A MOS n-channel cascode current mirror has a bias current of 0.3 mA and all transistor sizes are $W/L = 50\ \mu m/0.5\ \mu m$. What is the minimum output voltage allowable across the current mirror without any transistors entering the triode region? Use the NMOS device parameters for the 0.35-μm CMOS process in Table 1.5.

3.21 Using small-signal analysis, find the output resistance of a MOS cascode current mirror. Include in your analysis the voltage-dependent current source that models the body effect.

3.22 Shown in Fig. P3.22 is a PMOS current mirror with $V_{DD} = 1.8$ V, $I_{bias} = 150\ \mu A$, $R = 2\ k\Omega$, and all transistors sized $(W/L) = 18\ \mu m/0.2\ \mu m$ and having the device parameters for the 0.18-μm CMOS process in Table 1.5.

a. With $v_0 = 1$ V, are all transistors in active mode?

b. What is the maximum voltage that can appear at v_o while still keeping all transistors in active mode?

c. Assuming all transistors are in active mode, what is the small-signal output resistance seen looking into the drain of Q_4?

d. Compare this current mirror to the conventional cascode current mirror presented in Section 3.6.

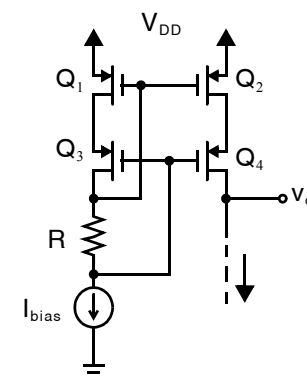


Fig. P3.22

3.11.7 Section 3.7: Cascode Gain Stage

3.23 A NMOS cascode amplifier with active load is shown in Fig. P3.23. It is realized using the 0.35-μm CMOS devices in Table 1.5 with $V_{DD} = 3.3$ V, $I_{bias} = 100\ \mu A$, $(W/L)_{1,2,5,6} = 10\ \mu m/0.5\ \mu m$, and $(W/L)_{3,4} = 30\ \mu m/0.5\ \mu m$.

a. Estimate the small-signal gain, v_o/v_i, and output resistance.

b. What is the total power consumption of the circuit?

c. Estimate the range of output voltages over which all transistors will remain in active mode.

3.24 Modify the design of Problem 3.23 in order to double its small-signal gain without changing its power consumption or the available output swing.

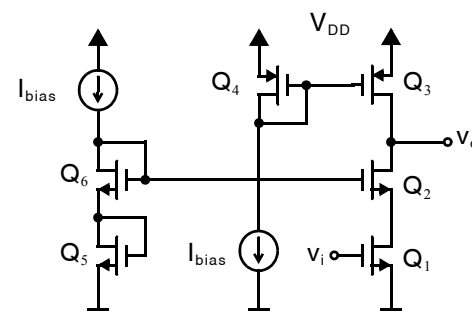


Fig. P3.23

3.25 Shown in Fig. P3.25 is another NMOS telescopic cascode amplifier, this time with the additional transistors $Q_{7,8}$ included. As in Problem 3.23, it is realized using the 0.35-μm CMOS devices in Table 1.5 with $V_{DD} = 3.3$ V, $I_{bias} = 100$ μA, $(W/L)_{1,2,5,6} = 10\ \mu m/0.5\ \mu m$, and $(W/L)_{3,4,7,8} = 30\ \mu m/0.5\ \mu m$.

a. Estimate the small-signal gain, v_o/v_i, and output resistance.

b. What is the total power consumption of the circuit?

c. Estimate the range of output voltages over which all transistors will remain in active mode.

d. How does this circuit compare with the amplifier in Problem 3.23?

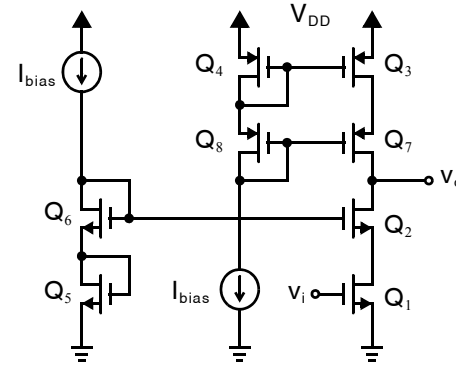


Fig. P3.25

3.26 A PMOS telescopic cascode amplifier is shown in Fig. P3.26. The transistor sizes, supply voltage and bias currents are the same as in Problem 3.25.

a. Estimate the small-signal gain, v_o/v_i, and output resistance.

b. Compare these results to the amplifier in Problem 3.25.

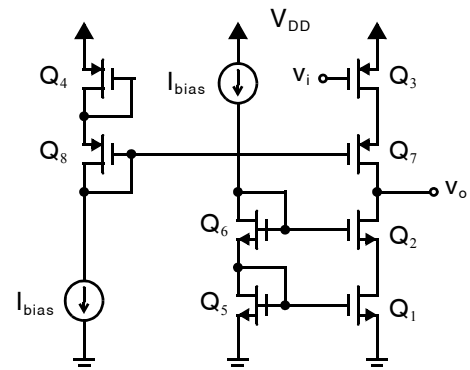


Fig. P3.26

3.11.8 Section 3.8: MOS Differential Pair and Gain Stage

3.27 For the differential-input stage in Fig. 3.19, assume that $I_{bias} = 150$ μA, all transistors have $W/L = 100\ \mu m/1.6\ \mu m$. Find the dc gain assuming the transistor parameters are those of the 0.8-μm CMOS parameters in Table 1.5.

3.28 In Fig. 3.19, $I_{bias} = 200$ μA, $(W/L)_{1,2} = 5\ \mu m/0.25\ \mu m$, and $(W/L)_{3,4} = 15\ \mu m/0.25\ \mu m$. The transistor parameters are those of the 0.18-μm CMOS parameters in Table 1.5 and the supply voltage is 1.8 V.

a. What is the output voltage, v_{out}, when the differential input voltage is zero? (You may assume all devices are in active mode.)

b. What is the minimum differential input voltage that will cause one of the input transistors $Q_{1,2}$ to turn off?

3.29 The CMOS differential pair circuit shown in Fig. P3.29 is to be designed to have a differential gain of 4 V/V and a power consumption of 1 mW. The transistors should be biased to have a $V_{eff} = 0.3$ V and

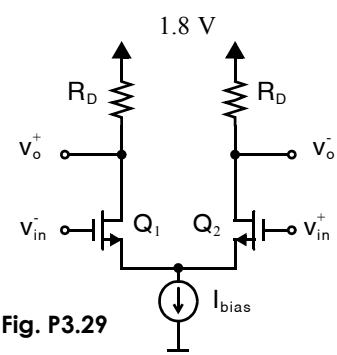


Fig. P3.29

should have a gate length of 0.25 μm. Using the 0.18-μm CMOS parameters in Table 1.5, determine the widths of the transistors and the values of R_D and I_{bias}. Compare your results with SPICE.

3.30 Replace the bias current I_{bias} in the circuit of Fig. P3.29 with the simple NMOS current mirror in Fig. 3.1. Select the device sizes, W and L, of the two current mirror transistors assuming $I_{in} = 25\ \mu A$. Ensure that the current mirror transistors remain in active mode for input voltages as low as 1 V.

3.31 Repeat Problem 3.29 using the 0.35-μm CMOS parameters in Table 1.5

3.32 The differential gain circuit in Fig. P3.32 is perfectly symmetric. Assume all transistors are kept in the active region.

a. Draw the equivalent half-circuit.

b. Derive an expression for the circuit's small-signal differential gain in terms of the transistor small-signal parameters: g_m, r_{ds}, etc.

c. Derive an expression for the circuit's small-signal output resistance in terms of the transistor small-signal parameters.

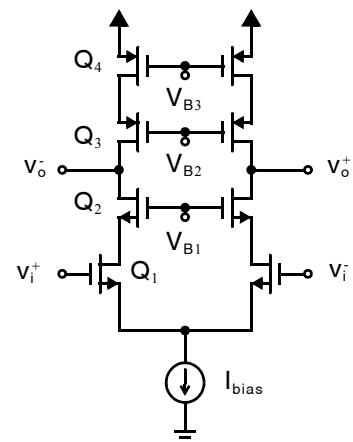


Fig. P3.32

3.33 The differential pair circuit shown in Fig. P3.33 is realized with transistors having the 0.18-μm CMOS parameters in Table 1.5. All gate lengths are $L = 0.3\ \mu m$, $I_{bias1} = 50\ \mu A$, and $I_{bias2} = 500\ \mu A$. The differential pair is perfectly symmetric. The device widths are related by $W_1 = W_2$ and $W_3 = W_4 = 4W_5 = 4W_6 = 4W_7$.

a. Select the NMOS device widths so that $V_{eff1,2} = 150$ mV.

b. Select the PMOS device widths so that $V_{eff3} = 300$ mV.

c. Estimate the differential small-signal gain of the circuit, $(v^+_{out} - v^-_{out})/(v^+_{in} - v^-_{in})$.

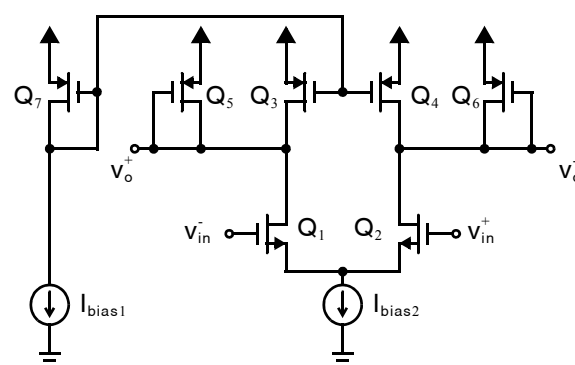


Fig. P3.33

CHAPTER

4 Frequency Response of Electronic Circuits

Historically, electronic circuits were characterized by exciting them with an oscillator's output, and then with an oscilloscope determining how the circuit affected the magnitude and phase of its input signal. In general, this technique is only appropriate for linear circuits. However, it is also useful for non-linear circuits containing transistors when the signals are small enough that the transistors can be adequately characterized by their operating points and small linear changes about their operating points; that is, the nonlinear transistors can be accurately described using small-signal analysis. The use of this technique for characterizing electronic circuits has led to the application of frequency-domain techniques to the analysis of most any linear or weakly non-linear system, a technique that is now ubiquitous in system analysis.

4.1 FREQUENCY RESPONSE OF LINEAR SYSTEMS

Consider a linear time-invariant system with transfer function $H(s)$ being excited by an input signal having Laplace transform $X_{in}(s)$.[1] The Laplace transform of the output signal, $X_{out}(s)$, is given by

$$X_{out}(s) = H(s)X_{in}(s) \tag{4.1}$$

In the time domain, assuming $x_{in}(t)$ is the inverse Laplace transform of $X_{in}(s)$, and $h(t)$ is the inverse Laplace Transform of $H(s)$ (often called its impulse response), we have

$$x_{out}(t) = x_{in}(t) \bullet h(t) = \int_{-\infty}^{\infty} x_{in}(\tau)h(t-\tau)d\tau \tag{4.2}$$

That is, the output signal is the convolution of the input signal with the impulse response of the system's transfer function.

The merit of frequency-domain analysis is that for particular inputs it is often easier to use (4.1) and the inverse Laplace Transform to calculate the expected output of a circuit rather than (4.2). Examples of this are when the inputs are pure sinusoids, or step inputs.

Only certain types of transfer functions arise in the ordinary course of analog circuit analysis. Since the inputs and outputs are generally voltages or currents in the circuit, x_{in} and x_{out} are always real-valued. Hence, the impulse response $h(t)$ is also always real-valued. Furthermore, we will restrict our discussion to circuits comprising lumped elements: resistors, capacitors, independent and dependent voltage and current sources.[2] The transfer

1. The Laplace transform of $x(t)$ is $X(s) = \int e^{-st}x(t)dt$.

2. We are excluding distributed circuit elements sometimes found in radio-frequency circuits such as transmission lines. These can give rise to irrational transfer functions.

functions of such circuits may always be written in rational form with real-valued coefficients. That is, as a ratio of polynomials in the Laplace Transform variable, s.

$$H(s) = \frac{a_0 + a_1 s + \ldots + a_m s^m}{1 + b_1 s + \ldots + b_n s^n} \tag{4.3}$$

In (4.3), all of the coefficients a_i and b_i are real. Normally, $m \le n$. When the transfer function is stable, all of the b_i will be positive.

Transfer functions may also be written as a ratio of products of first-order terms,

$$\begin{aligned} H(s) &= K\frac{(s+z_1)(s+z_2)\ldots(s+z_m)}{(s+\omega_1)(s+\omega_2)\ldots(s+\omega_n)} \\ &= a_0\frac{\left(1+\frac{s}{z_1}\right)\left(1+\frac{s}{z_2}\right)\ldots\left(1+\frac{s}{z_m}\right)}{\left(1+\frac{s}{\omega_1}\right)\left(1+\frac{s}{\omega_2}\right)\ldots\left(1+\frac{s}{\omega_n}\right)} \end{aligned} \tag{4.4}$$

where $K = a_m/b_n$. These are often referred to as *root form* since the roots of the numerator (*zeros*) and denominator (*poles*) are clearly visible. However, to be exact note that z_i and ω_i are not the actual roots of the numerator and denominator; rather they are equal to the negative roots. Hence, the transfer functions zeros are $-z_1, -z_2\ldots$ and the poles are $-\omega_1, -\omega_2\ldots$. It is also common to refer to the *frequencies* of the zeros and poles which are always positive quantities, $|z_1|, |z_2|\ldots$ and $|\omega_1|, |\omega_2|\ldots$, since these are physical frequencies of significance in describing the system's input-output behavior.

> ***Key Point:*** *The transfer functions in analog circuits throughout this text:*
> *a) are rational with* $m \le n$
> *b) have real-valued coefficients,* a_i *and* b_i
> *c) have poles and zeros that are either real or appear in complex-conjugate pairs*
> *Furthermore, if the system is stable,*
> *d) all denominator coefficients* $b_i > 0$
> *e) the real part of all poles will be negative*

For transfer functions of systems with real-valued inputs and outputs, all zeros and poles will be either real or occur in complex-conjugate pairs. Further, the real parts of all the poles will be positive for stable transfer functions.

4.1.1 Magnitude and Phase Response

The output of an oscillator can often be characterized as a sinusoidal signal given by

$$x_{in}(t) = A_{in}\cos(\omega_{in}t) = A_{in}\frac{(e^{j\omega_{in}t} + e^{-j\omega_{in}t})}{2} \tag{4.5}$$

where Euler's relation has been used

$$e^{jx} = \cos(x) + j\sin(x) \tag{4.6}$$

and

$$j = \sqrt{-1} \tag{4.7}$$

is a purely imaginary number of unit magnitude.

In the frequency domain, $X_{in}(s)$ consists of two impulses, one at

$$s = j\omega_{in} \tag{4.8}$$

and one at

$$s = -j\omega_{in} \tag{4.9}$$

with each impulse having a magnitude $A_{in}/2$.

The output of a linear system having such an input also consists of two impulses in the frequency domain, but now each impulse will be multiplied by $H(s)$ where $s = j\omega_{in}$ for the positive frequency impulse and $s = -j\omega_{in}$ for the negative frequency impulse. After taking the inverse Laplace Transform of the output signal, it is seen that

$$x_{out}(t) = \frac{A_{in}}{2}(e^{j\omega_{in}t}H(j\omega_{in}) + e^{-j\omega_{in}t}H(-j\omega_{in})) \tag{4.10}$$

The transfer function evaluated on the imaginary axis, $H(j\omega_{in})$, is its *frequency response* which may written in terms of its magnitude response $|H(j\omega_{in})|$ and phase response $\phi = \angle H(j\omega_{in})$.

$$H(j\omega_{in}) = |H(j\omega_{in})|e^{j\phi} \tag{4.11}$$

Furthermore, for systems with real-valued inputs and outputs,

$$H(-j\omega_{in}) = |H(j\omega_{in})|e^{-j\phi} \tag{4.12}$$

Substituting (4.11) and (4.12) into (4.10) gives

$$\begin{aligned} x_{out}(t) &= \frac{A_{in}}{2}|H(j\omega_{in})|(e^{j\omega_{in}t}e^{j\phi} + e^{-j\omega_{in}t}e^{-j\phi}) \\ &= \frac{A_{in}}{2}|H(j\omega_{in})|(e^{j(\omega_{in}t+\phi)} + e^{-j(\omega_{in}t+\phi)}) \\ &= A_{in}|H(j\omega_{in})|\cos(\omega_{in}t + \phi) \end{aligned} \tag{4.13}$$

The development just presented has been given in terms of the Laplace transform variable s for the particular case where $s = j\omega_{in}$. Alternatively, it is often presented in terms of Fourier Transforms which are equivalent to Laplace Transforms with $s = j\omega_{in}$. However, when using Fourier Transforms, the j symbol denoting that the frequency is complex is usually omitted. For example, (4.11) would normally be written as

$$H(\omega_{in}) = |H(\omega_{in})|e^{j\phi} \tag{4.14}$$

where $\phi = \angle H(\omega_{in})$.

> ***Key Point:*** *When a linear circuit is excited with a sinusoid, the output will be a sinusoid at the same frequency. Its magnitude will equal to the input magnitude multiplied by the magnitude response* $|H(\omega_{in})|$. *The phase difference between the output and input sinusoids will equal the phase response* $\angle H(\omega_{in})$. *Magnitude responses are often expressed in units of decibels,* $20\log_{10}|H(\omega)|$ dB.

The interpretation of (4.14) is as follows: When a linear electronic circuit is excited with the output of a sinusoidal oscillator, the output of the circuit will also be a sinusoid at the same frequency. The magnitude of the output sinusoid will be equal to the magnitude of the input sinusoid multiplied by $|H(\omega_{in})|$ where $H(\omega_{in})$ is the frequency response of the circuit. The phase difference between the output sinusoid and the input sinusoid will be equal to $\phi = \angle H(\omega_{in})$.

Magnitude response is often expressed in units of decibels (dB),

$$20\log_{10}|H(\omega)| \text{ dB}$$

Decibels are a convenient unit since the magnitude response of two linear systems in series is the sum of the two magnitude responses when expressed in dB. Specifically, two systems with frequency responses $H_1(\omega)$ and $H_2(\omega)$ connected in series will have an overall frequency response $H_1(\omega) \cdot H_2(\omega)$. Expressed in decibels,

$$20\log_{10}(|H_1(\omega)||H_2(\omega)|) \text{ dB} = 20\log_{10}|H_1(\omega)| \text{ dB} + 20\log_{10}|H_2(\omega)| \text{ dB}$$

EXAMPLE 4.1

If the magnitude response of a linear system is doubled, how much does it increase in dB? What if the magnitude response is increased by an order of magnitude?

Solution

If the magnitude response of the original system is $20\log_{10}|H(\omega)|$ dB, once it is doubled it becomes

$$\begin{aligned} 20\log_{10}(2|H(\omega)|) \text{ dB} &= 20\log_{10}|H(\omega)| \text{ dB} + 20\log_{10}2 \text{ dB} \\ &= 20\log_{10}|H(\omega)| \text{ dB} + 6.02 \text{ dB} \\ &\cong 20\log_{10}|H(\omega)| \text{ dB} + 6 \text{ dB} \end{aligned}$$

If increased by an order of magnitude, it becomes

$$\begin{aligned} 20\log_{10}(10|H(\omega)|) \text{ dB} &= 20\log_{10}|H(\omega)| \text{ dB} + 20\log_{10}10 \text{ dB} \\ &= 20\log_{10}|H(\omega)| \text{ dB} + 20 \text{ dB} \end{aligned}$$

Hence, doubling a magnitude response implies an increase of 6 dB and increasing it by one order of magnitude implies an increase of 20 dB. In general, a change in $|H(\omega)|$ by a multiplicative factor $|A|$ can be expressed in decibels by adding $20\log_{10}|A|$ dB .

4.1.2 First-Order Circuits

A first-order transfer function has a first order denominator, $n = 1$. For example,

$$H(s) = \frac{A_0}{1 + \dfrac{s}{\omega_0}} \tag{4.15}$$

is a first-order low-pass transfer function. It is the most-commonly encountered transfer function in electronic circuits. It arises naturally when a resistance and capacitance are combined. It is also often used as a simple model of more complex circuits, such as operational amplifiers; when used in this way, ω_0 is referred to as a *dominant pole* and (4.15) is a *dominant-pole approximation*.

EXAMPLE 4.2

Consider a linear circuit having a transfer function given by (4.15), where $A_0 = 10$ and $\omega_0 = 2\pi \times 100$ rad./s. Find the magnitude and phase of the output sinusoid assuming the input sinusoid has a peak voltage of 1V and zero phase for the frequencies 10Hz, 100Hz, and 1kHz.

Solution

We can re-write (4.15) with $s = j\omega$ using Fourier Transforms as

$$H(\omega) = \frac{A_0}{1 + j\dfrac{\omega}{\omega_0}} = \frac{A_0}{\sqrt{1 + \left(\dfrac{\omega}{\omega_0}\right)^2}} e^{j\phi} \tag{4.16}$$

where

$$\phi = -\tan^{-1}\left(\frac{\omega}{\omega_0}\right) \tag{4.17}$$

Note that

$$|H(\omega)| = \frac{A_0}{\sqrt{1 + \left(\dfrac{\omega}{\omega_0}\right)^2}} \tag{4.18}$$

Using (4.17) and (4.18), for the case $f = 10$ Hz, where $\omega = 2\pi f = 628.3$ rad, we have

$$|H(\omega)| = \frac{10}{\sqrt{1 + \left(\dfrac{2\pi f}{2\pi f_0}\right)^2}} = \frac{10}{\sqrt{1 + \left(\dfrac{f}{f_0}\right)^2}} = \frac{10}{\sqrt{1 + \left(\dfrac{10}{100}\right)^2}} = 9.95 = 19.96 \text{ dB} \tag{4.19}$$

The phase shift is given by

$$\phi = -\tan^{-1}\left(\frac{10}{100}\right) = -5.7° \tag{4.20}$$

The fact that the phase difference between the input and output is negative, implies the output sinusoid comes after the input sinusoid by –5.7°, or in other words "lags" the input sinusoid. For $f_{in} = 100$ Hz, we have

$$|H(\omega)| = \frac{10}{\sqrt{2}} = 0.7071 \times 10 = 20 \text{ dB} - 3.01 \text{ dB} \tag{4.21}$$

and

$$\phi = -\tan^{-1}(1) = -45° \tag{4.22}$$

Note that at $f_{in} = f_0$ or equivalently, for $\omega_{in} = \omega_0$, the magnitude response is 3dB below its low-frequency value, and the phase is delayed by –45°. Finally, for $f_{in} = 1$kHz, we have

$$H(\omega)| = \frac{10}{\sqrt{1 + \left(\dfrac{f}{f_0}\right)^2}} = \frac{10}{\sqrt{1 + \left(\dfrac{1000}{100}\right)^2}} = \frac{10}{10.05} = 20 \text{ dB} - 20.04 \text{ dB} = -0.04 \text{ dB} \tag{4.23}$$

and

$$\phi = -\tan^{-1}\left(\frac{1000}{10}\right) = -1.47\text{rad} = -84° \tag{4.24}$$

Thus, the gain is approximately 20dB less than its low-frequency value and the phase shift is approaching –90°.

Step Response of First-Order Circuits

Another common means of characterizing linear circuits is to excite them with step inputs. This would normally be done in practice by exciting them with a square wave having a low-enough frequency (or equivalently long-enough period), so the circuit settles between edges, and therefore each edge effectively excites the circuit similar to a step input.

Consider a step input $x_{in}(t) = A_{in}u(t)$ where $u(t)$, the step-input function, is defined as

$$u(t) = \begin{cases} 0, & t \le 0 \\ 1, & t > 0 \end{cases} \tag{4.25}$$

The step function $u(t)$ has a Laplace transform given by

$$U(s) = \frac{1}{s} \tag{4.26}$$

Therefore, the Laplace transform of the output of a linear system having transfer function $H(s)$ and a step input is given by

$$X_{out}(s) = A_{in}\frac{H(s)}{s} \tag{4.27}$$

This is normally easy to calculate especially since any rational $H(s)$ may be expressed as a sum of first-order terms using *residues* [James, 2004].

Consider the special case were a first-order low-pass filter has a transfer function given by

$$H(s) = \frac{A_0}{1 + \dfrac{s}{\omega_0}} \tag{4.28}$$

Using (4.27), we have

$$X_{out}(s) = \frac{A_{in}}{s}\frac{A_0}{1 + \dfrac{s}{\omega_0}} \tag{4.29}$$

Equation (4.29) can be rewritten in terms of its residues as

$$X_{out}(s) = A_{in}A_0\left[\frac{1}{s} - \frac{1}{s + \omega_0}\right] \tag{4.30}$$

It is now simple to take the Inverse Laplace Transform to get

$$x_{out}(t) = u(t)A_{in}A_0[1 - e^{-t/\tau}] \tag{4.31}$$

where $\tau = 1/\omega_0$.

In a similar manner it is straight-forward to show that a general first-order circuit having a transfer function given by

$$H(s) = A_0\left(\frac{1+\frac{s}{\omega_z}}{1+\frac{s}{\omega_0}}\right) \tag{4.32}$$

and subjected to a step input has an output with Laplace transform

$$X_{out}(s) = \frac{A_{in}A_0}{s}\left(\frac{1+\frac{s}{\omega_z}}{1+\frac{s}{\omega_0}}\right) \tag{4.33}$$

Hence, the system's step response is

$$x_{out}(t) = u(t)A_{in}A_0\left[1-\left(1-\frac{\omega_0}{\omega_z}\right)e^{-t/\tau}\right] \tag{4.34}$$

where $\tau = 1/\omega_0$

Note that the step response for t very slightly greater than 0 is denoted by $x_{out}(0^+)$ and using (4.34) is easily found to be given by

$$x_{out}(0^+) = A_{in}A_0\frac{\omega_0}{\omega_z} \tag{4.35}$$

This is easily verified using the Laplace transform property

$$x_{out}(0^+) = \lim_{s\to\infty} sX_{out}(s) = \lim_{s\to\infty} H(s)A_{in} = A_{in}A_0\frac{\omega_0}{\omega_z} \tag{4.36}$$

together with equation (4.33). In a similar manner, it is easily found that after a long time, when the first-order circuit has settled, its response denoted $x_{out}(\infty)$ is given by

$$x_{out}(\infty) = A_{in}A_0 \tag{4.37}$$

Once again, this can be verified using the Laplace Transform property

$$x_{out}(\infty) = \lim_{s\to 0} sX_{out}(s) = \lim_{s\to 0} H(s)A_{in} \tag{4.38}$$

together with (4.33). Using (4.34), (4.35) and (4.38), one can derive the general equation giving the step response of any first-order circuit as

$$x(t) = x(\infty) - [x(\infty) - x(0^+)]e^{-t/\tau} \tag{4.39}$$

EXAMPLE 4.3

Consider the first-order lowpass circuit shown in Fig. 4.1. Assume $R = 1\ k\Omega$ and $C = 1$ nF. Find the –3 dB frequency of the circuit and the time constant of the circuit. Assuming the input signal is a 0.5-V step input at time 0, what is the output voltage at 1.5 μs? How long does it take for the circuit to settle within 70% of its final value?

Solution

The transfer function of this circuit is easily found using the admittance divider formula; that is

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{G_1(s)}{G_1(s) + G_2(s)} \tag{4.40}$$

where $G_1(s)$ is the interconnecting admittance between the input and output and the denominator is the sum of all admittances connected to the output node (in this case $G_1(s) + G_2(s)$). For this case, we have $G_1(s) = 1/R$ and $G_2(s) = sC$ and therefore

$$H(s) = \frac{1/R}{sC + 1/R} = \frac{1}{1 + sRC} \tag{4.41}$$

The –3dB frequency in radians per second is given by

$$\omega_0 = \frac{1}{RC} = \frac{1}{1\times10^3 1\times10^{-9}} = 1\times10^6 \text{ rad/sec} \tag{4.42}$$

In Hz, the –3dB frequency is given by

$$f_{-3dB} = \frac{1}{2\pi}\frac{1}{RC} \cong 159 \text{ kHz} \tag{4.43}$$

The circuit time constant is given by

$$\tau = \frac{1}{\omega_0} = RC = 1\ \mu s \tag{4.44}$$

Next, using (4.31), we have the output voltage at 1.5 μs given by

$$x_{out}(t) = 0.5[1 - e^{-1.5\times10^{-6}/1\times10^{-6}}] = 0.5[1 - e^{-1.5}] = 0.39V \tag{4.45}$$

For the circuit to settle within 70% of its final value, we need

$$1 - e^{-t/\tau} = 0.7 \tag{4.46}$$

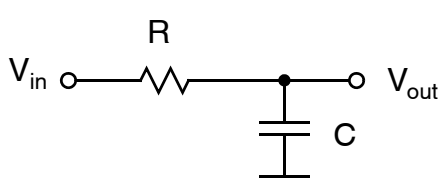


Fig. 4.1 A RC first-order lowpass circuit.

which implies

$$t = \tau \ln\left(\frac{1}{1-0.7}\right) = 1.2\tau = 1.2\mu s \tag{4.47}$$

EXAMPLE 4.4

Consider the first-order circuit shown in Fig. 4.2. Find equations for the transfer function and for the step-response. What constraint is necessary and sufficient for the step-response to be an ideal step? For the special case of $C_1 = 5pF$, $C_2 = 10pF$, $R_1 = 2k\Omega$, and $R_2 = 10k\Omega$, plot the step response assuming the input signal is a 2V step.

Solution

The transfer function can once again be found using the admittance-divider formula. We have

$$\begin{aligned} H(s) &= \frac{V_{out}(s)}{V_{in}(s)} = \frac{sC_1 + 1/R_1}{sC_1 + 1/R_1 + sC_2 + 1/R_2} \\ &= \left(\frac{R_2}{R_1 + R_2}\right)\left[\frac{1 + sR_1C_1}{1 + s\dfrac{(R_1R_2)}{(R_1 + R_2)}(C_1 + C_2)}\right] \end{aligned} \tag{4.48}$$

Equation (4.48) can be re-written in the same form as (4.32) with the substitutions

$$A_0 = \frac{R_2}{R_1 + R_2} \tag{4.49}$$

$$\tau_z = \omega_z^{-1} = R_1C_1 \tag{4.50}$$

and

$$\tau_0 = \omega_0^{-1} = \frac{(R_1R_2)}{(R_1 + R_2)}(C_1 + C_2) = (R_1 \parallel R_2)(C_1 + C_2) \tag{4.51}$$

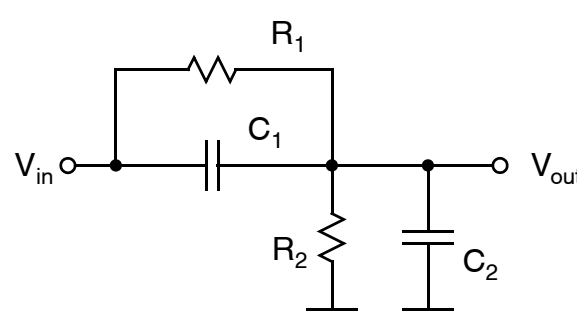


Fig. 4.2 A RC first-order circuit.

In order for the step response (given by (4.34)) to be an ideal step, we need $\omega_z = \omega_0$. Using (4.50) and (4.51), we see the necessary conditions are

$$R_1C_1 = (R_1 \parallel R_2)(C_1 + C_2) \tag{4.52}$$

which implies

$$\frac{C_2}{C_1} = \frac{R_1}{R_2} \tag{4.53}$$

For the case, of $C_1 = 5\text{pF}$, $C_2 = 10\text{pF}$, $R_1 = 2\text{k}\Omega$, and $R_2 = 10\text{k}\Omega$, we have $A_0 = 0.833$, $\omega_z = 1\times10^8\,\text{rad}$, and $\omega_0 = 4\times10^7\,\text{rad}$. Using (4.36) and (4.48), we see

$$x_{out}(0^+) = A_{in}\lim_{s\to\infty}H(s) = A_{in}\frac{C_1}{C_1 + C_2} = 2\left(\frac{5\text{pF}}{5\text{pF} + 10\text{pF}}\right) = 0.67\text{V} \tag{4.54}$$

and

$$x_{out}(\infty) = A_{in}\lim_{s\to 0}H(s) = A_{in}\frac{R_2}{R_1 + R_2} = 2\left(\frac{10\text{k}\Omega}{2\text{k}\Omega + 10\text{k}\Omega}\right) = 1.67\text{V} \tag{4.55}$$

The circuit time constant is $1/\omega_0 = 25\text{ns}$. Using these values, it is possible to sketch the step response as shown in Fig. 4.3. In one time constant, the output voltage has settled to within 22% of its total change (i.e. $x_{out}(\infty) - (x_{out}(\infty) - x_{out}(0^+))e^{-1} = 1.67\text{V} - 1\text{V}\cdot e^{-1} = 1.30\text{V}$).

EXAMPLE 4.5

Consider an amplifier having a small signal transfer function that approximately is given by

$$A(s) = \frac{A_0}{1 + \dfrac{s}{\omega_{-3\text{dB}}}} \tag{4.56}$$

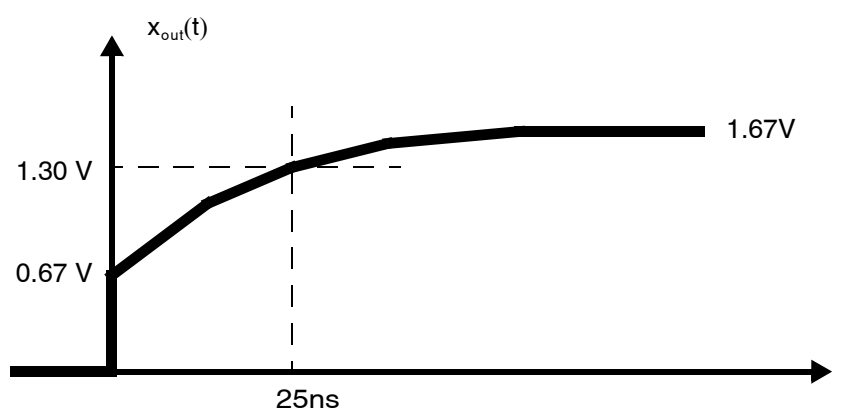


Fig. 4.3 The 2-V step-response of the circuit of Fig. 4.2.

What is the approximate unity-gain frequency and the phase shift at the unity-gain frequency for $A_0 = 1\times10^5$ and $\omega_{-3dB} = 1\times10^3$?

Solution

At frequencies $s = j\omega \gg j\omega_{-3dB}$, which is the case for most signal frequencies, the s term in the denominator dominates and we have

$$A(s) \cong \frac{A_0\omega_{-3dB}}{s} \tag{4.57}$$

and the open-loop transfer function of the amplifier is very similar to that of an ideal integrator. The magnitude of the Fourier Transform is found by substituting $s = j\omega$ and taking the magnitude. To find the unity gain frequency, this magnitude should be set to unity and then ω (often called ω_{ta}) should be solved for. We have

$$\left|\frac{A_0\omega_{-3dB}}{\omega_{ta}}\right| = 1 \tag{4.58}$$

> ***Key Point:*** *For a first-order lowpass transfer function with dc gain* $A_0>>1$*, the unity gain frequency is* $\omega_{ta}\approx A_0\omega_{-3dB}$ *and* $\angle A(\omega_{ta})\approx 90°$.

which implies

$$\omega_{ta} \cong A_0\omega_{-3dB} \tag{4.59}$$

The phase shift of the transfer function is given by

$$\angle A(\omega) = -\tan\left(\frac{\omega}{\omega_{-3dB}}\right)^{-1} \tag{4.60}$$

Noting that since $A_0 \gg 1$, we have $\omega/\omega_{-3dB} \gg 1$, and therefore

$$\angle A(\omega_{ta}) \cong -90° \tag{4.61}$$

For this example, using (4.59) and (4.61), we have $\omega_{ta} \cong 1\times10^8$ rad.

4.1.3 Second-Order Low-Pass Transfer Functions with Real Poles

Second-order low-pass transfer functions are often encountered when characterizing electronic circuits. For example, operational amplifiers with feedback, and perhaps being inadequately compensated, are often modelled as second-order low-pass functions. Three alternative formulations for second-order low-pass transfer functions (assuming a real numerator) are

$$\begin{aligned} H(s) &= \frac{K}{(1+s\tau_1)(1+s\tau_2)} \\ &= \frac{K}{\left(1+\frac{s}{\omega_{p1}}\right)\left(1+\frac{s}{\omega_{p2}}\right)} \\ &= \frac{K\omega_{p1}\omega_{p2}}{(s+\omega_{p1})(s+\omega_{p2})} \end{aligned} \tag{4.62}$$

The coefficients τ_1, τ_2 or ω_{p1}, ω_{p2} are either real and positive, or occur in complex-conjugate pairs. The transfer function may also be written in the form

$$H(s) = \frac{K\omega_{p1}\omega_{p2}}{\omega_{p1}\omega_{p2} + s(\omega_{p1} + \omega_{p2}) + s^2} \tag{4.63}$$

Equation (4.63) may be written in the alternative (and popular) form

$$H(s) = \frac{K\omega_0^2}{\omega_0^2 + s\frac{\omega_0}{Q} + s^2} \tag{4.64}$$

Here, parameter ω_o is called the resonant frequency, parameter Q is called the Q-factor[3] [Sedra, 2009], and K is the dc gain of the transfer function. This latter form (4.64) is often preferred by experienced designers because it has no complex-valued coefficients, even in cases where the poles $-\omega_{p1}$ and $-\omega_{p2}$ are a complex-conjugate pair.

We will for now consider only the special case where $\omega_{p1} < \omega_{p2}$ are both real and distinct. We can set the denominator of (4.64), $D(s)$, equal to the denominator of (4.62), and we have

$$\begin{aligned} D(s) &= \omega_0^2 + s\frac{\omega_0}{Q} + s^2 \\ &= (\omega_{p1} + s)(\omega_{p2} + s) \\ &= \omega_{p1}\omega_{p2} + s(\omega_{p1} + \omega_{p2}) + s^2 \end{aligned} \tag{4.65}$$

Equating the coefficients yields

$$\omega_0^2 = \omega_{p1}\omega_{p2} \tag{4.66}$$

and

$$\frac{\omega_0}{Q} = \omega_{p1} + \omega_{p2} \tag{4.67}$$

Solving (4.66) and (4.67), we get

$$\omega_{p1}, \omega_{p2} = \frac{\omega_0}{2Q}(1 \pm \sqrt{1 - 4Q^2}) \tag{4.68}$$

We can also express the frequency response of (4.62) as

$$H(s) = \frac{K}{\sqrt{1 + \left(\frac{\omega}{\omega_{p1}}\right)^2}\sqrt{1 + \left(\frac{\omega}{\omega_{p2}}\right)^2}} e^{j\phi} \tag{4.69}$$

where

$$\phi = -\tan^{-1}\left(\frac{\omega}{\omega_{p1}}\right) - \tan\left(\frac{\omega}{\omega_{p2}}\right) \tag{4.70}$$

This form clearly shows $|H(\omega)|$ and $\angle H(\omega)$.

3. The Q-factor is 1/2 times the inverse of the damping-factor. The damping-factor is an alternative method of indicating the pole locations in second-order transfer-functions.

Widely-Spaced Real Poles

For the special case of real poles $\omega_{p1} \ll \omega_{p2}$, we have $Q \ll 1$, and we can make the approximation

$$\sqrt{1-4Q^2} \cong 1-2Q^2 \tag{4.71}$$

Using (4.71) and (4.68) gives

$$\omega_{p1} \cong \omega_0 Q \tag{4.72}$$

$$\omega_{p2} \cong \frac{\omega_0}{Q} \tag{4.73}$$

Using (4.69) and (4.70), it is possible to simply approximate the magnitude and phase response of low-pass transfer functions in a number of different frequency regions. For $\omega \ll \omega_p$, we have $|H(\omega)| = K$. At $\omega = \omega_{p1}$, we have $H(\omega_{p1}) = K/\sqrt{2}$, and $\angle H(\omega) = -45°$. Next, for $\omega_{p1} \ll \omega \ll \omega_{p2}$, we have

$$|H(\omega)| \cong \frac{K\omega_{p1}}{\omega} \tag{4.74}$$

If we express the magnitude response in terms of decibels (dB),

$$|H(\omega)|_{dB} \cong 20\log(K\omega_{p1}) - 20\log(\omega) \tag{4.75}$$

Thus, the magnitude response is decreasing –20dB for every decade increase in ω. In addition, in this region, we have $\angle H(\omega) \cong -90°$. Next, at $\omega = \omega_{p2}$, we have

$$|H(\omega_{p2})| \cong \frac{K\omega_{p1}}{\sqrt{2}\omega_{p2}} \tag{4.76}$$

and $\angle H(\omega) = -135°$. The final region is when $\omega_{p2} \ll \omega$. In this region, we have

$$|H(\omega)| \cong \frac{K\omega_{p1}\omega_{p2}}{\omega^2} \tag{4.77}$$

Expressing the magnitude gain in dB, we have

$$|H(\omega)|_{dB} \cong 20\log(K\omega_{p1}\omega_{p2}) - 40\log(\omega) \tag{4.78}$$

and we see the gain is decreasing –40 dB for every decade increase in ω. In this region, we also have $\angle H(\omega) \cong -180°$.

Step Response

The step response for a second-order low-pass transfer function can be found by expanding (4.62) into partial fractions (or residues). It is not too difficult to show that

$$H(s) = K\left[\frac{\omega_{p2}}{\omega_{p2}-\omega_{p1}}\frac{\omega_{p1}}{s+\omega_{p1}} - \frac{\omega_{p1}}{\omega_{p2}-\omega_{p1}}\frac{\omega_{p2}}{s+\omega_{p2}}\right] \tag{4.79}$$

Equivalently, in terms of time constants $\tau_1 = 1/\omega_{p1}$ and $\tau_2 = 1/\omega_{p2}$

$$H(s) = K\left[\frac{\tau_1}{\tau_1-\tau_2}\frac{1}{1+s\tau_1} - \frac{\tau_2}{\tau_1-\tau_2}\frac{1}{1+s\tau_2}\right] \tag{4.80}$$

It is now straightforward to express the step response of the second-order system as the sum of the individual step responses of two first-order systems. We have

$$x_{out}(t) = A_{in}\frac{K\omega_{p2}}{\omega_{p2}-\omega_{p1}}[1-e^{-t\omega_{p1}}] - A_{in}\frac{K\omega_{p1}}{\omega_{p2}-\omega_{p1}}[1-e^{-t\omega_{p2}}] \tag{4.81}$$

and we see the step response is composed of two first-order modes.

For the case of widely-spaced real poles $\omega_{p1} \ll \omega_{p2}$, we see that the second term in (4.81) settles much more quickly then the first. Hence, for $t \gg 1/\omega_{p2}$,

$$x_{out}(t) \cong A_{in}\frac{K\omega_{p2}}{\omega_{p2}-\omega_{p1}}[1-e^{-t\omega_{p1}}] - A_{in}\frac{K\omega_{p1}}{\omega_{p2}-\omega_{p1}} \cong A_{in}K[1-e^{-t\omega_{p1}}] \tag{4.82}$$

The system exhibits a first-order step response with the slower time constant, $1/\omega_{p1}$.

4.1.4 Bode Plots

The observations made in Section 4.1.3 are the basis of a methodology useful for constructing approximate plots of the magnitude response (in dB) and phase response (in degrees) versus log frequency. These plots are called *Bode Plots*; they are used extensively to graphically represent a circuit's frequency response and for the analysis of feedback systems.

The method for Bode plots of stable low-pass transfer functions having only real poles (i.e. where the order of the numerator $m = 0$, hence $N(s) = K$) is summarized by the following rules:

a. At very low frequencies both the magnitude and phase plots are constant (horizontal lines) at $20\log_{10}|H(0)|$ dB and $0°$ respectively.

b. As the frequency becomes larger than a pole frequency, the slope of the magnitude response changes by –20 dB/decade.

c. Each pole contributes an additional $-45°$ phase shift at the frequency of the pole. At frequencies substantially higher than a pole frequency, it contributes an additional $-90°$.

Key Point: *Transfer functions with only real-valued poles and numerator order zero have monotonically decreasing magnitude and phase response plots with each pole contributing an additional –20 dB/decade slope in the magnitude response, and an additional –90° phase shift.*

EXAMPLE 4.6

Sketch a bode plot for a second-order low-pass transfer function (4.62) with $K = 1\times10^4$, $\omega_{p1} = 10$ rad/s, and $\omega_{p2} = 1000$ rad/s.

Solution

The transfer function is

$$H(s) = \frac{1\times10^4}{\left(1+\frac{s}{10}\right)\left(1+\frac{s}{1000}\right)}$$

Following the procedure, the magnitude response is constant for low frequencies at a value of $20\log_{10}|H(0)| = 20\log_{10}10^4 = 80$ dB (step a). The slope then decreases to –20 dB/decade at $\omega_{p1} = 10$ rad/s and to –40 dB/decade at $\omega_{p2} = 10^3$ rad/s (step b). The phase response starts at $0°$ (step a) and decreases to $-45°$ at $\omega_{p1} = 10$, $-90°$ at $\omega_{p1} \ll \omega \ll \omega_{p2}$, to $-135°$ at $\omega_{p2} = 10^3$ rad/s, and finally to $-180°$ at $\omega \gg \omega_{p2}$ (step c). The plots in Fig. 4.4 result.

The methodology described above may be generalized to handle any rational transfer function where all zeros and poles are purely real. It is easiest to have $H(s)$ written in root form, (4.4).

a. Find a frequency ω_{mid} where the number of poles at lower frequencies, $\omega_1, \omega_2 \ldots \omega_k$, equals the number of zeros at lower frequencies, $z_1, z_2 \ldots z_k$. Begin the magnitude response plot with a flat (constant) region around ω_{mid} at a value of

$$20\log_{10}|H(\omega_{mid})| \cong 20\log_{10}\left(K\frac{|z_{(k+1)}|\cdots|z_m|}{|\omega_{(k+1)}|\cdots|\omega_n|}\right)$$

Begin the phase response plot at $\angle H(\omega_{mid}) \cong 0°$.

b. Moving from ω_{mid} to the right along the magnitude response plot (to higher frequencies), each time the frequency becomes greater than a pole frequency, ω_i with $i > k$, the slope of the magnitude response changes by –20 dB/decade. Each time the frequency becomes greater than a zero frequency, $|z_i|$ with $i > k$, the slope of the magnitude response changes by +20 dB/decade.

c. Moving from ω_{mid} to the left along the magnitude response plot (to lower frequencies), each time the frequency becomes less than a pole frequency, ω_i with $i \le k$, the slope of the magnitude response changes by +20 dB/decade. Each time the frequency becomes less than a zero frequency, $|z_i|$ with $i \le k$, the slope of the magnitude response changes by –20 dB/decade.

d. Moving from ω_{mid} to the right along the phase response plot (to higher frequencies), an additional $-45°$ phase shift is introduced at the frequency of each pole, ω_i with $i > k$, and right-half plane zero, $|z_i|$ with $i > k$ and $z_i < 0$; at frequencies substantially higher than the pole and zero frequencies an additional $-90°$ phase shift is contributed. At each left-half-plane zero encountered, $|z_i|$ with $i > k$ and $z_i > 0$, an

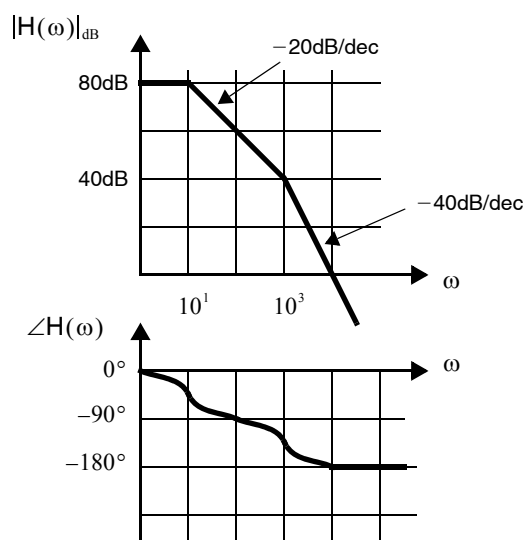


Fig. 4.4 Magnitude gain (in dB) and phase versus ω on a logarithmic scale.

additional +45° phase shift is introduced at the frequency $|z_i|$, and an additional +90° phase shift is contributed at frequencies substantially higher.

e. Moving from ω_{mid} to the left along the phase response plot (to lower frequencies), an additional +45° phase shift is introduced at the frequency of each pole, ω_i with $i \le k$, and right-half plane zero, $|z_i|$ with $i \le k$ and $z_i < 0$; at frequencies substantially lower than the pole and zero frequencies an additional +90° phase shift is contributed. At each left-half-plane zero encountered, $|z_i|$ with $i \le k$ and $z_i > 0$, an additional −45° phase shift is introduced at the frequency $|z_i|$, and an additional −90° phase shift is contributed at frequencies substantially lower.

EXAMPLE 4.7

Find the transfer function of the opamp circuit shown in Fig. 4.5 and sketch its Bode plots.

Solution

Since C_1 is five orders of magnitude larger than C_2, there is a wide range of frequencies where C_1 can be assumed to be a short circuit and C_2 can be approximated by an open circuit.[4] In this frequency range, called midband, we have simply a noninverting opamp configuration with a gain of $(1 + R_2/R_1)$.

At moderately-low frequencies where C_1 is having an effect, C_2 can certainly be treated as an open circuit since it is so small. The circuit at these frequencies appears like the simplified circuit shown in Fig. 4.6. The transfer function at these frequencies is found by analyzing this simplified circuit. The closed-loop transfer function is given by

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = 1 + \frac{Z_2(s)}{Z_1(s)} \tag{4.83}$$

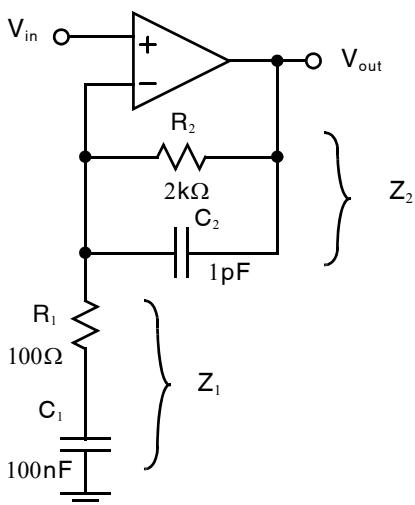


Fig. 4.5 The feedback configuration often used for stereo power amplifiers.

4. The reason for including C_1 is that at very-very-low frequencies, it operates as an open circuit. Thus, at dc, the gain is only unity. The low gain at d.c. (of unity) helps minimize bothersome "clicks" due to d.c. offsets inherent in the amplifier that would be amplified by the large gain if C_1 had not been included. The capacitor C_2 limits the high frequency response. It also helps stabilize the opamp, but this is beyond the scope of the current subject.

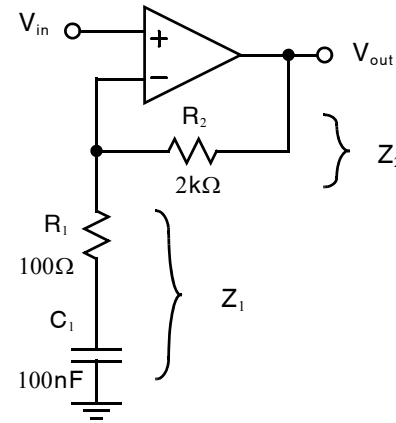


Fig. 4.6 A simplified circuit that has the same frequency response as Fig. 4.5 for low and mid-band frequencies.

where

$$Z_1(s) = R_1 + \frac{1}{sC_1} \tag{4.84}$$

and

$$Z_2(s) = R_2 \tag{4.85}$$

We have

$$H(s) = \frac{R_1 + R_2 + \dfrac{1}{sC_1}}{R_1 + \dfrac{1}{sC_1}} = \frac{1 + s(R_1 + R_2)C_1}{1 + sR_1C_1} \tag{4.86}$$

We see the simplified circuit is first order with a zero at

$$\omega_{z1} = \frac{1}{(R_1 + R_2)C_1} = 4.76 \text{ krad/sec} \tag{4.87}$$

and a pole at

$$\omega_{p1} = \frac{1}{R_1C_1} = 100 \text{ krad/sec} \tag{4.88}$$

Note that as the frequency gets large, $\omega \gg \omega_{z1}, \omega_{p1}$, the result in (4.86) becomes the same as that obtained earlier by inspection for mid-band frequencies.

$$H(s) \cong 1 + \frac{R_2}{R_1} \tag{4.89}$$

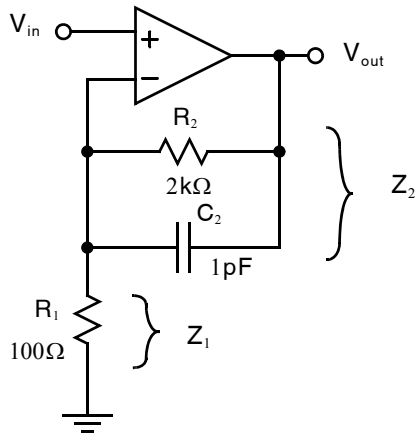


Fig. 4.7 A simplified circuit that has the same frequency response as that in Fig. 4.5 for mid-band and high frequencies.

At mid and high-frequencies, when C_1 can be certainly be treated as a short-circuit, the response of the circuit is very close to that of the simplified circuit shown in Fig. 4.7. We now have

$$Z_1(s) = R_1 \tag{4.90}$$

and

$$Z_2(s) = \frac{1}{1/R_2 + sC_2} \tag{4.91}$$

Using (4.83), we have (after a little algebra)

$$\begin{aligned} H(s) &= \frac{Y_1(s) + Y_2(s)}{Y_2(s)} \\ &= \frac{\frac{1}{R_1} + \frac{1}{R_2} + sC_2}{\frac{1}{R_2} + sC_2} \\ &= \left(\frac{R_1 + R_2}{R_1}\right)\left(\frac{1 + s(R_1 \parallel R_2)C_2}{1 + sR_2C_2}\right) \end{aligned} \tag{4.92}$$

Key Point: *When a circuit has one or more capacitors with relatively very large value, one may consider those capacitors short-circuited to see how the circuit behaves at higher frequencies. Similarly, when one or more capacitors have relatively very small value, the circuit may be considered with those capacitors open-circuited to understand its operation at lower frequencies.*

We see the simplified circuit is again first order with a zero at

$$\omega_{z2} = \frac{1}{(R_1 \parallel R_2)C_2} = 1.05\times10^8 \text{ rad/sec} \tag{4.93}$$

and a pole at

$$\omega_{p2} = \frac{1}{R_2C_2} = 5\times10^6 \text{ rad/sec} \tag{4.94}$$

Once again the mid-band gain may be found by evaluating (4.92) at low frequencies, $\omega \ll \omega_{z2}, \omega_{p2}$ resulting in $H(s) \cong 1 + (R_2/R_1)$.

An alternative approach would be to consider the actual circuit of Fig. 4.5 that is valid at all frequencies and substitute

$$Z_1(s) = R_1 + \frac{1}{sC_1} \tag{4.95}$$

and

$$Z_2(s) = \frac{1}{1/R_2 + sC_2} \tag{4.96}$$

directly into (4.83) without simplifying. After some algebra, we get

$$H(s) = \frac{1 + s[(R_1 + R_2)C_1 + R_2C_2] + s^2R_1R_2C_1C_2}{(1 + sR_1C_1)(1 + sR_2C_2)} \tag{4.97}$$

If we now make the assumption that the numerator, $N(s)$, has two widely separated zeros $\omega_{z1} \ll \omega_{z2}$, then we may express the numerator as

$$\begin{aligned} N(s) &= \left(1 + \frac{s}{\omega_{z1}}\right)\left(1 + \frac{s}{\omega_{z2}}\right) \\ &= 1 + s\left(\frac{1}{\omega_{z1}} + \frac{1}{\omega_{z2}}\right) + \frac{s^2}{\omega_{z1}\omega_{z2}} \\ &\cong 1 + \frac{s}{\omega_{z1}} + \frac{s^2}{\omega_{z1}\omega_{z2}} \end{aligned} \tag{4.98}$$

Equating (4.98) and (4.97) and solving for ω_{z1} and ω_{z2}, we get

$$\omega_{z1} \cong \frac{1}{(R_1 + R_2)C_1 + R_2C_2} \cong \frac{1}{(R_1 + R_2)C_1} \tag{4.99}$$

$$\omega_{z2} \cong \frac{1}{\omega_{z1}R_1R_2C_1C_2} \cong \frac{R_1 + R_2}{R_1R_2C_2} = \frac{1}{(R_1 \parallel R_2)C_2} \tag{4.100}$$

We also have two widely-spaced poles given by

$$\omega_{p1} = \frac{1}{R_1C_1} \tag{4.101}$$

$$\omega_{p2} = \frac{1}{R_2C_2} \tag{4.102}$$

Hence, we have the same poles and zeros that were obtained by separate analysis of the low-frequency circuit in Fig. 4.6 and the high-frequency circuit in Fig. 4.7.

Following the procedure for sketching Bode plots, we begin at mid-band frequencies $\omega_{p1} \ll \omega \ll \omega_{p2}$. Here, the gain as given by (4.89) is 21 and the magnitude of the bode plot is constant at 26.4 dB while the phase is constant at 0°. Moving to higher frequencies, we first encounter the pole ω_{p2} and then the zero ω_{z2}. Moving to lower frequencies, we first encounter the pole ω_{p1} and then the zero ω_{z1}. The resulting Bode plot is shown in Fig. 4.8.

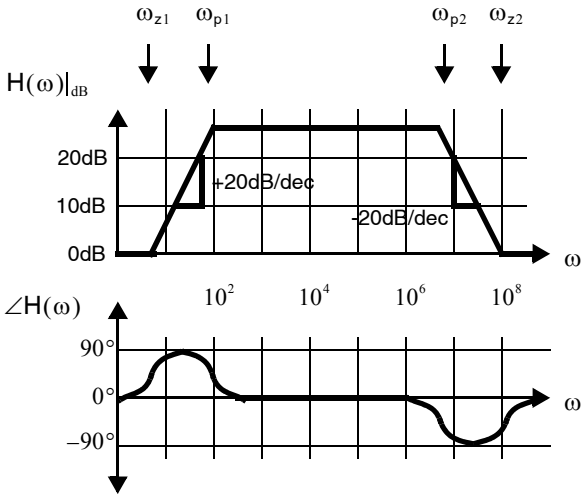


Fig. 4.8 Approximate magnitude and phase responses of the circuit of Fig. 4.5.

4.1.5 Second-Order Low-Pass Transfer Functions with Complex Poles

Consider the second-order low-pass transfer function in (4.62) when the two poles are complex-valued and conjugate pairs, $\omega_{p1} = \omega_{p2}^*$. Using (4.68), we see this case occurs when $Q > 1/2$ and we have

$$\omega_{p1}, \omega_{p2} = \frac{\omega_0}{2Q}[1 \pm j\sqrt{4Q^2 - 1}] \equiv \omega_r \pm j\omega_q \tag{4.103}$$

Hence,

$$\omega_r = \frac{\omega_0}{2Q} \tag{4.104}$$

and

$$\omega_q = \frac{\omega_0}{2Q}\sqrt{4Q^2 - 1} \tag{4.105}$$

Furthermore,

$$|\omega_{p1}| = |\omega_{p2}| = \frac{\omega_0}{2Q}[1 + 4Q^2 - 1]^{1/2} = \omega_0 \tag{4.106}$$

The pole locations on the complex s-plane are illustrated in Fig. 4.9. The step response can now be found by substituting (4.103) into (4.81).

$$\begin{aligned} x_{out}(t) &= \frac{A_{in}K}{-2j\omega_q}[(\omega_r - j\omega_q)(1 - e^{-\omega_r t}e^{-j\omega_q t}) - (\omega_r + j\omega_q)e^{-\omega_r t}e^{j\omega_q t}] \\ &= \frac{A_{in}K}{2j\omega_q}(j\omega_q[2 - e^{-\omega_r t}(e^{j\omega_q t} + e^{-j\omega_q t})] - \omega_r[e^{-\omega_r t}(e^{j\omega_q t} - e^{-j\omega_q t})]) \\ &= A_{in}K\left[1 - e^{-\omega_r t}\cos(\omega_q t) - \frac{\omega_r}{\omega_q}e^{-\omega_r t}\sin(\omega_q t)\right] \\ &= A_{in}K[1 - A_s e^{-\omega_r t}\cos(\omega_q t + \theta)] \end{aligned} \tag{4.107}$$

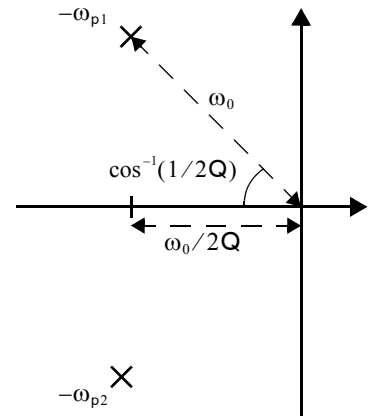


Fig. 4.9 The complex-conjugate pole locations of a second-order stable transfer function with $Q > 0.5$.

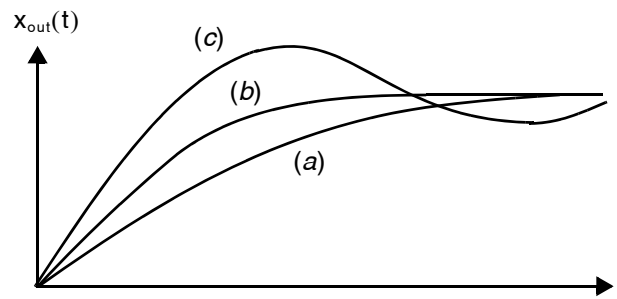


Fig. 4.10 The step responses of a second-order low-pass circuit for (*a*) $Q < 0.5$ (*b*) $Q = 0.5$, and (*c*) $Q > 0.5$

where $A_s = \sqrt{2-4Q^2}$ and $\theta = \tan^{-1}\sqrt{4Q^2-1}$. Thus, we see the step response has a sinusoidal term whose envelope exponentially decreases with a time constant equal to the inverse of the real parts of the poles, $1/\omega_r = 2Q/\omega_0$. Hence, a system with high Q-factor will exhibit oscillations that persist for a long time. The oscillatory frequency of the sinusoids are determined by the imaginary parts of the poles.

It is also possible to determine the peak of the step response by solving

$$\frac{d}{dt}x_{out}(t) = 0 \tag{4.108}$$

for time, t. In general, there may be multiple solutions to (4.108), but since the envelope of the oscillations is decaying, the solution with the smallest magnitude $|t|$ is the time of the highest peak in the step response. The height is obtained by evaluating (4.107) at this time,

$$x_{out}(t)\big|_{max} = A_{in}Ke^{-\left(\frac{\pi}{\sqrt{4Q^2-1}}\right)} \tag{4.109}$$

Equation (4.109) is valid only for $Q > 0.5$. For $Q \le 0.5$, the poles are real-valued and there is no overshoot. The borderline case $Q = 0.5$ is called a maximally-damped response. These different cases have been plotted approximately in Fig. 4.10.

4.2 FREQUENCY RESPONSE OF ELEMENTARY TRANSISTOR CIRCUITS

When analyzing electronic circuits containing transistors to determine their frequency response, a small-signal analysis is implicitly assumed since only linearized systems can have a well defined frequency response. Given this assumption, the procedure is the same one outlined at the very beginning of Chapter 3, except that the circuit's parasitic capacitances are included to capture the circuit's high-frequency limitations.

4.2.1 High-Frequency MOS Small-Signal Model

The small-signal high-frequency model for MOS transistors takes the low-frequency model and adds parasitic capacitances to model the charge-storage effects of MOS transistors. These effects were covered in detail in Chapter 1, but the salient points are repeated here in sufficient detail to permit basic frequency analysis of MOS transistor circuits.

A cross-sectional view of a MOS transistor showing the parasitic capacitances is shown in Fig. 4.11. The high-frequency model for MOS transistors in the active region is shown in Fig. 4.12. The largest capacitance in the model is the gate-source capacitance C_{gs}. For a transistor with width W and gate length L, its value can be estimated with the approximate formula

$$C_{gs} = \frac{2}{3}WLC_{ox} + C_{ov} \tag{4.110}$$

where C_{ov} is the gate to source or drain overlap capacitance and C_{ox} is the gate capacitance per unit area. The capacitance C_{gd} is primarily due to the overlap capacitance,

$$C_{gd} = C_{ov} \tag{4.111}$$

The other capacitances in the small-signal model are due to the depletion capacitances of reverse-biased junctions. We have

$$C_{sb} = (A_s + WL)C_{js} + P_sC_{j\text{-}sw} \tag{4.112}$$

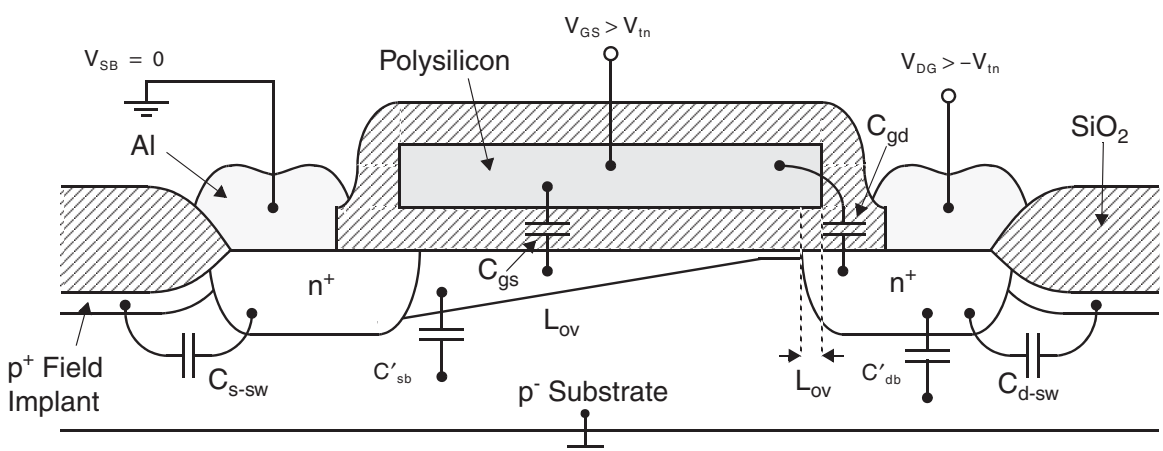


Fig. 4.11 A cross-section of an n-channel MOS transistor showing the small signal capacitances.

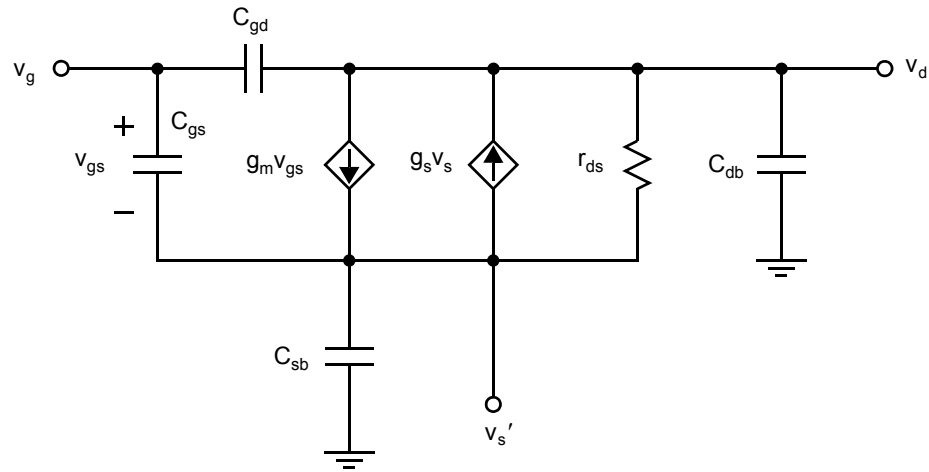


Fig. 4.12 The high-frequency model for MOS transistors in the active region.

where A_s is the area of the source junction, P_s is the effective perimeter of the source junction, and C_{js} and $C_{j\text{-}sw}$ are the per-unit-area and per-unit-length capacitances formed underneath and around the side-walls of the source junction region. The effective area used for the source junction in (4.112) has the channel area added (i.e. WL) as the source junction and the channel are electrically connected with low impedance. The equation for the drain-bulk capacitance is similar.

$$C_{db} = A_d C_{jd} + P_d C_{j\text{-}sw} \tag{4.113}$$

The effective drain junction does not include the area of the channel (i.e. WL). Both C_{js} and C_{jd} are inversely proportional to $\sqrt{1 + V_{SB \text{ or } DB}/\Phi_0}$. This accounts for the increase in depletion region thickness with increasing reverse bias on the source and drain body junctions. Since $V_{DB} > V_{SB}$ in active mode, C_{jd} will be slightly less than C_{js}.

4.2.2 Common-Source Amplifier

A common-source amplifier is shown in Fig. 4.13. It is here driven by a voltage source v_{in} with source resistance R_s, although this may of course be the Thevenin equivalent of a input current source having the same source

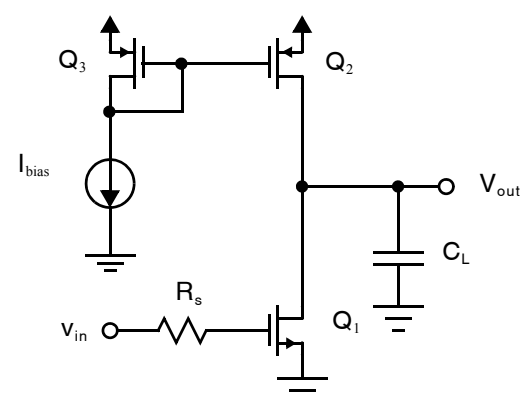


Fig. 4.13 A common-source amplifier.

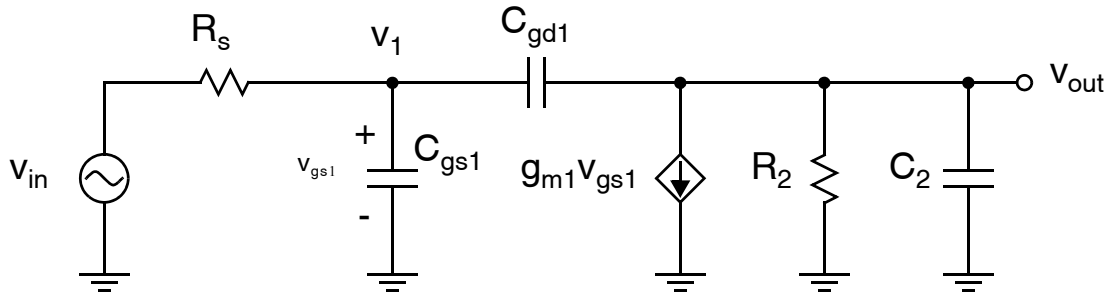


Fig. 4.14 A small-signal model for high-frequency analysis of the common-source amplifier.

resistance. It is assumed that the bias voltage of the input signal is such that both Q_1 and Q_2 are in the active region. Based on this assumption, a small-signal equivalent circuit for high-frequency analysis of the common-source amplifier of Fig. 4.13 is shown in Fig. 4.14. Here, C_{gs1} is the gate-to-source capacitance of Q_1 and C_{gd1} is the gate-to-drain capacitance of Q_1. Note that it has been assumed that the output capacitance of the input signal source can be ignored. The capacitance C_2 is made up of the parallel connection of the drain-to-bulk capacitances of Q_1 and Q_2, the gate-drain capacitance of Q_2, and the load capacitance C_L.

$$C_2 = C_{db1} + C_{db2} + C_{gd2} + C_L \tag{4.114}$$

Similarly, R_2 is the parallel combination of r_{ds1} and r_{ds2}, both small-signal resistances between the output node and small-signal ground.

$$R_2 = r_{ds1} \| r_{ds2} \tag{4.115}$$

The frequency response of the small-signal circuit may be obtained by nodal analysis. To avoid making circuit equation errors, a consistent methodology should be maintained when writing nodal equations. Here, the first term is always the node at which the currents are being summed. Multiplying this node voltage is the sum of all admittances connected to the node. The next negative terms are the adjacent node voltages with each being multiplied by the connecting admittance. The last terms are any current sources with a positive sign being used if the current sources flow out of the node. Thus, the nodal equation at v_1 is

$$v_1(G_s + sC_{gs1} + sC_{gd1}) - v_{in}G_s - v_{out}sC_{gd1} = 0 \tag{4.116}$$

where $G_s = 1/R_s$[5]. Also, at the output node, we have

$$v_{out}(G_2 + sC_{gd1} + sC_2) - v_1 sC_{gd1} + g_{m1}v_1 = 0 \tag{4.117}$$

where we have used $v_1 = v_{gs1}$. From (4.117), we have

$$v_1 = v_{out}\frac{[G_2 + s(C_{gd1} + C_2)]}{-g_{m1} + sC_{gd1}} \tag{4.118}$$

Substituting (4.118) into (4.116) gives

$$v_{out}\frac{[G_2 + s(C_{gd1} + C_2)][G_s + s(C_{gs1} + C_{gd1})]}{-g_{m1} + sC_{gd1}} - v_{out}sC_{gd1} = v_{in}G_s \tag{4.119}$$

5. Whenever a variable is designated G_i, it is implicitly assumed that it is an admittance and that $G_i = 1/R_i$, where R_i is the resistance of the same component.

which implies

$$v_{out}\left\{\begin{array}{c} G_sG_2 + s[G_2(C_{gs1} + C_{gd1}) + G_s(C_{gd1} + C_2) + g_{m1}C_{gd1}] \\ + \; s^2[(C_{gs1} + C_{gd1})(C_{gd1} + C_2) - C_{gd1}^2] \end{array}\right\} = -v_{in}G_s(g_{m1} - sC_{gd1}) \tag{4.120}$$

Continuing, we have

$$A(s) = \frac{v_{out}}{v_{in}} = \frac{-g_{m1}R_2\left(1 - s\dfrac{C_{gd1}}{g_{m1}}\right)}{1 + sa + s^2b} \tag{4.121}$$

where

$$a = R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2) \tag{4.122}$$

and

$$b = R_sR_2(C_{gd1}C_{gs1} + C_{gs1}C_2 + C_{gd1}C_2) \tag{4.123}$$

This result may seem surprisingly complex for a single-transistor circuit, but several important features are revealed by careful consideration of (4.121) – (4.123).

The low frequency gain is obtained by substituting $s = 0$ into (4.121) and is, as expected,

$$A_0 = -g_{m1}R_2 \tag{4.124}$$

Assuming the poles are real and widely separated with $\omega_{p1} \ll \omega_{p2}$, the denominator can be expressed as

$$D(s) = \left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{p2}}\right) \cong 1 + \frac{s}{\omega_{p1}} + \frac{s^2}{\omega_{p1}\omega_{p2}} \tag{4.125}$$

Equating the coefficients of (4.125) with those of the denominator of (4.121), we have

$$\omega_{p1} \cong \frac{1}{a} = \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)} \tag{4.126}$$

$$\omega_{p2} \cong \frac{1}{\omega_{p1}b} \tag{4.127}$$

Substituting (4.123) and (4.126) into (4.127), we have

$$\omega_{p2} \cong \frac{g_{m1}C_{gd1}}{C_{gs1}C_{gd1} + C_{gs1}C_2 + C_{gd1}C_2} \tag{4.128}$$

Finally, at high frequencies the zero can be important.

$$\omega_z = -\frac{g_{m1}}{C_{gd1}} \tag{4.129}$$

Intuitively, the zero is caused because at high frequencies C_{gd1} shorts the gate and drain of Q_1 together, providing a direct path from the amplifier's input to output. Hence, as frequency increases beyond $|\omega_z|$, the circuit appears to have one less node and one less pole. As a result the roll-off in magnitude response provided by the poles is stifled and the slope of $|A(\omega)|_{dB}$ increases by +20 dB/decade. Because the sign of ω_z is negative, the zero is in the right

half-plane and therefore causes phase lag rather than phase lead. This is important when compensating CMOS opamps having common-source stages.

Since $\omega_{p1} \ll \omega_{p2}, \omega_z$, a dominant-pole approximation may be applied for frequencies $\omega \ll \omega_{p2}, \omega_z$,

$$A(s) \cong \frac{A_0}{1 + \dfrac{s}{\omega_{p1}}} = \frac{-g_m R_2}{1 + s\{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)\}} \tag{4.130}$$

and the 3-dB bandwidth is approximately the dominant pole frequency.

$$\omega_{-3dB} \cong \omega_{p1} = \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)} \tag{4.131}$$

For the special case where the load capacitance and therefore C_2 is large, we have

$$\omega_{p1} \cong \frac{1}{R_2 C_2} \tag{4.132}$$

$$\omega_{p2} \cong \frac{C_{gd1}}{C_{gs1} + C_{gd1}} \frac{g_{m1}}{C_2} \tag{4.133}$$

Note that both ω_{p2} and ω_z are proportional to the transistor's transconductance g_{m1}; having a large transconductance is important in minimizing the detrimental effects of the second pole and the zero by pushing them to higher frequencies.

The high-frequency analysis of the common-source amplifier illustrates the critical importance of utilizing approximations in analog circuits, both to expedite analysis and to build intuition. The next two subsections present the two most important approximations invoked for the high-frequency analysis of analog circuits, and apply them to the common-source amplifier.

Key Point: *The frequency-response of the common-source amplifier has 2 poles and 1 zero. The complexity of the analysis illustrates the importance of making approximations and simplifications in the high-frequency analysis of analog circuits.*

4.2.3 Miller Theorem and Miller Effect

The most commonly presented method for analyzing common-source amplifiers to determine their -3dB frequency is based on the Miller theorem. This theorem states that the two circuits shown in Fig. 4.15 are equivalent if Y_1 and Y_2 are chosen appropriately. Specifically, if

$$Y_1(s) = Y(s)(1 + A) \tag{4.134}$$

and

$$Y_2(s) = Y(s)\left(1 + \frac{1}{A}\right) \tag{4.135}$$

then the current through Y_1 is

$$\begin{aligned} I_1 &= Y_1(s)V_1 \\ &= Y(s)(1 + A)V_1 \\ &= Y(s)(V_1 + AV_1) \\ &= Y(s)(V_1 - V_2) \\ &= I \end{aligned}$$

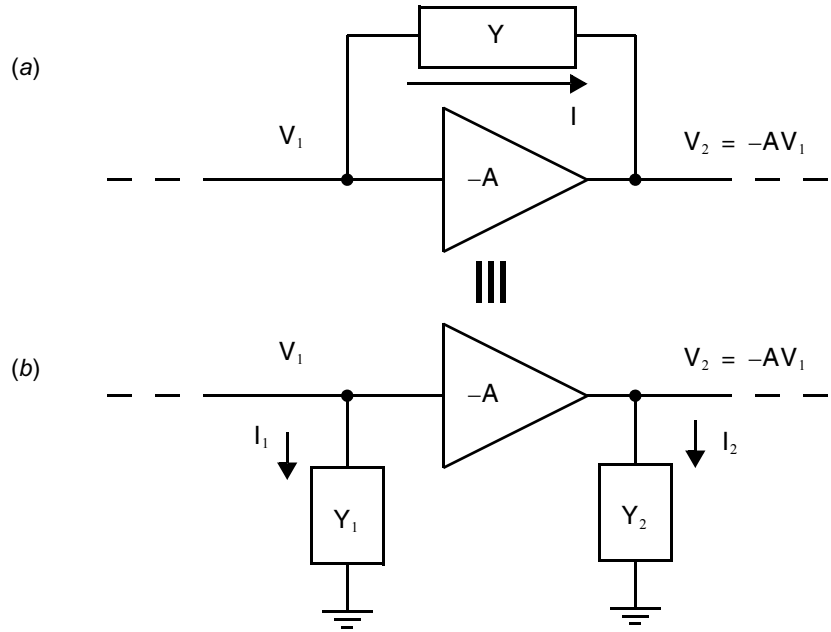


Fig. 4.15 The Miller theorem states that circuit (*a*) and (*b*) are equivalent assuming Y_1 and Y_2 are chosen appropriately.

and the current through Y_2 is

$$\begin{aligned} I_2 &= Y_2(s)V_2 \\ &= Y(s)\left(1 + \frac{1}{A}\right)V_2 \\ &= Y(s)\left(V_2 + \frac{1}{A}V_2\right) \\ &= Y(s)(V_2 - V_1) \\ &= -I \end{aligned}$$

Hence, in both circuits the currents leaving nodes V_1 and V_2 through the impedance branches are the same, so all nodal equations are identical and analysis of the two circuits will produce the same results.

EXAMPLE 4.8

What is the input impedance of an ideal voltage amplifier with gain $-A$ and a capacitor C connecting the input and output?

Solution

This is a straightforward application of Miller Theorem to the case where $Y(s) = 1/sC$. Additional insight is gained by recognizing that any capacitance C may be represented by two capacitors in series, as shown in Fig. 4.16.

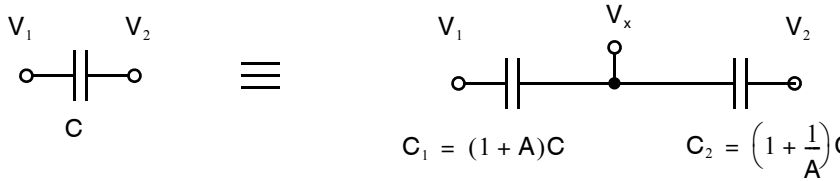


Fig. 4.16 Equivalence of a capacitance C and two series capacitors, C_1 and C_2.

$$C_1 = (1 + A)C \tag{4.136}$$

$$C_2 = \left(1 + \frac{1}{A}\right)C \tag{4.137}$$

This is true for any constant A since

$$\frac{C_1 C_2}{C_1 + C_2} = \frac{(1 + A)C \cdot \left(1 + \frac{1}{A}\right)C}{(1 + A)C + \left(1 + \frac{1}{A}\right)C} = \frac{\left(A + 2 + \frac{1}{A}\right)C^2}{\left(A + 2 + \frac{1}{A}\right)C} = C$$

When C is split into C_1 and C_2 as shown in Fig. 4.16, a new (fictitious) node is created at V_x whose potential is a simple voltage division between V_1 and V_2.

$$V_x = V_1 + \frac{C_2}{C_1 + C_2}(V_2 - V_1) \tag{4.138}$$

In this case, C is connected around an ideal voltage amplifier with gain $-A$, so $V_2 = -AV_1$. Substituting this, along with (4.136) and (4.137) into (4.138) gives the interesting result,

$$V_x = V_1 + \left(\frac{\left(1 + \frac{1}{A}\right)C}{(1 + A)C + \left(1 + \frac{1}{A}\right)C}\right)(-AV_1 - V_1)$$

$$= V_1 - \left(\frac{\left(1 + \frac{1}{A}\right)(A + 1)C}{(1 + A)C + \left(1 + \frac{1}{A}\right)C}\right)V_1 = 0$$

Hence, choosing C_1 and C_2 as in (4.136) and (4.137) provides a perfect voltage-division ratio to ensure that the node V_x remains at zero potential (i.e. ground) at all times. The equivalency is illustrated in Fig. 4.17. A similar procedure may be applied to any complex admittance $Y(s)$, providing an alternate proof of the Miller Theorem.

Clearly, the input impedance Z_{in} is the capacitance $C_1 = (1 + A)C$, a much larger value than the actual capacitor C when the gain $|A|$ is large. This amplification of C by the gain of an inverting amplifier is called the *Miller effect*, and C is the *Miller capacitor*.

Intuitively, the Miller effect may be understood as follows. When the voltage at the amplifier input changes by ΔV_1, the voltage across the Miller capacitor changes by $(1 + A)\Delta V_1$. Thus, whatever source is driving V_1 must deliver a charge $(1 + A)\Delta V_1 C$ to the Miller capacitor—the same charge that would be required by a *grounded* capacitor with the much larger value $(1 + A)C$.

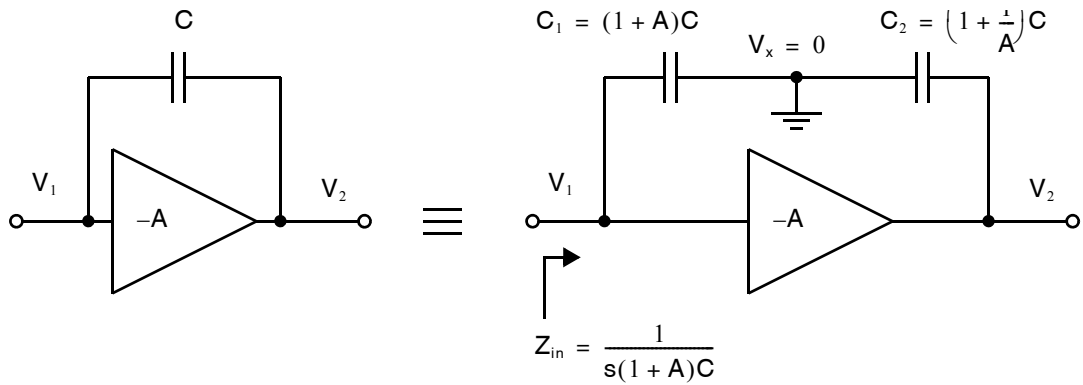


Fig. 4.17 The application of the Miller theorem to a capacitance across an ideal voltage amplifier results in a large effective input capacitance.

Key Point: *When a capacitor connects the input and output of a high-gain inverting amplifier, it appears much larger at the input than it really is. This Miller effect is often responsible for the dominant pole of such amplifiers and, hence, is useful for obtaining a quick estimate of bandwidth.*

The power of the Miller theorem is that it greatly simplifies the analysis for the dominant pole of a circuit having a capacitor between the input and output of an inverting high-gain amplifier, as we shall see. However, the Miller theorem is not used when estimating the second pole of amplifiers. At frequencies where capacitive loading at node V_2, which is responsible for the second pole, becomes appreciable, the gain between nodes V_1 and V_2 starts to decrease. Hence, $|A(\omega)|$ may no longer be assumed constant and application of the Miller theorem is difficult.

The Miller theorem can be used to simplify the analysis of the small-signal circuit of Fig. 4.14. The simplified circuit is shown in Fig. 4.18. This simplification is based on noting that the low-frequency gain between v_1 and v_{out} is

$$\frac{v_{out}}{v_1} \equiv -A = -g_{m1}R_2 \tag{4.139}$$

The admittance Y bridging v_1 and v_{out} is $Y(s) = sC_{gd1}$. Therefore, we have

$$Y_1(s) = sC_{gd1}(1 + A) = sC_{gd1}(1 + g_{m1}R_2) \tag{4.140}$$

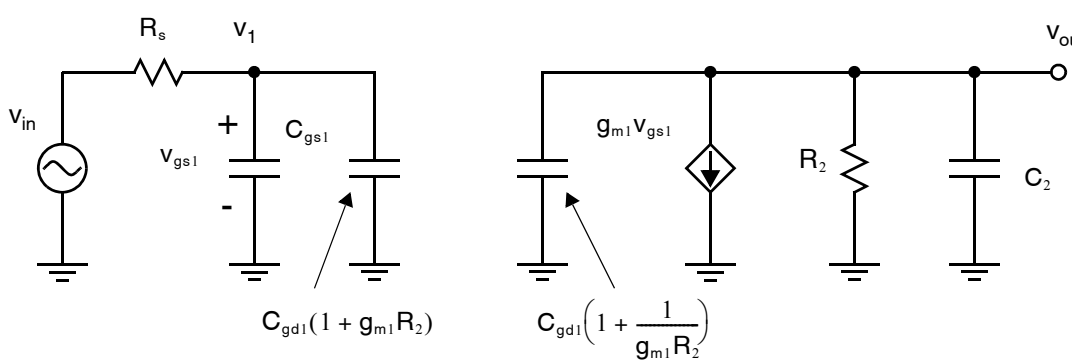


Fig. 4.18 A simplified small-signal model for the common-source amplifier applying the Miller theorem.

and

$$Y_2(s) = sC_{gd1}\left(1 + \frac{1}{A}\right) = sC_{gd1}\left(1 + \frac{1}{g_{m1}R_2}\right) \tag{4.141}$$

It is now evident that the transfer function from v_{in} to v_1 has a low-frequency pole which certainly dominates over any others because of the large capacitance there.

$$\frac{v_1}{v_{in}}(s) \cong \frac{1}{1 + \frac{s}{\omega_{p1}}} \tag{4.142}$$

where

$$\omega_{p1} = \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)]} \tag{4.143}$$

The estimate for the –3dB frequency of the circuit is then given by

$$\omega_{-3dB} \cong \omega_{p1} = \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)]} \tag{4.144}$$

This is in agreement with the result obtained by nodal analysis in (4.131) except that terms related to the output node, $R_2(C_{gd1} + C_2)$, are here ignored. Because the size of C_{gd1} is effectively multiplied by one plus the gain of the amplifier, the importance of having C_{gd1} small is obvious.

As a final note regarding the Miller theorem, this technique of realizing a large grounded capacitor by placing a smaller capacitor between the input and output of a high-gain inverting amplifier may be used to effectively realize large grounded on-chip capacitors.

4.2.4 Zero-Value Time-Constant Analysis

The most common and powerful technique for the high-frequency analysis of complex circuits is the zero-value time-constant analysis technique [Gray, 2009]. It is very powerful in estimating a circuit's bandwidth with minimal complication and also in determining which nodes are most important and need to have their associated time-constants minimized. Generally, the approach is to calculate a time-constant for each capacitor in the circuit by assuming all other capacitors are zero, then sum all the time constants to provide a single overall time constant. Beginning with the small-signal equivalent circuit, the procedure is as follows.

a. Set all independent sources to zero. That is, make all voltage sources into short circuits and all current sources into open circuits.

b. For each capacitor C_k in turn, with all other capacitors taken to be zero (making them open circuits), find a corresponding time-constant. To do this, replace the capacitor in question with a voltage source, and then calculate the resistance "seen by that capacitor," R_k, by taking the ratio of the voltage source to the current flowing from it. Note that in this analysis step, there are no capacitors in the circuit. The corresponding time-constant is then simply the capacitor multiplied by the resistance it sees: $\tau_k = R_kC_k$.

Key Point: *The method of zero-value time-constants estimates the dominant pole and, hence, bandwidth of complex circuits by analyzing several simpler circuits at dc.*

c. The –3dB frequency for the complete circuit is approximated by one over the sum of the individual capacitor time-constants.[6]

$$\omega_{-3dB} \cong \frac{1}{\sum \tau_k} = \frac{1}{\sum R_k C_k} \tag{4.145}$$

This procedure is best understood by way of example.

EXAMPLE 4.9

Perform a zero-value time-constant analysis on the common-source amplifier.

Solution

Begin with the small-signal equivalent circuit in Fig. 4.14. Set $v_{in} = 0$. There are three capacitances in the circuit: C_{gs1}, C_2, and C_{gd1}. We shall consider each in turn.

To compute the time-constant for C_{gs1}, set $C_2 = 0$ and $C_{gd1} = 0$ making them open circuits. Then replace C_{gs1} with a test voltage source, v_1. The resulting small-signal circuit, shown in Fig. 4.19(*a*), must then be analyzed to find $R_1 = v_1/i_1$. The analysis is trivial, yielding $R_1 = R_s$ and the first time constant,

$$\tau_1 = R_s C_{gs1} \tag{4.146}$$

For the capacitor C_2, open-circuit both C_{gs1} and C_{gd1} and apply a test voltage v_2 at the output resulting in the small-signal circuit in Fig. 4.19(*b*). Note that there is no current in R_s so that $v_{gs1} = 0$ and $g_{m1}v_{gs1} = 0$. The result follows straightforwardly:

$$\tau_2 = R_2 C_2 \tag{4.147}$$

The analysis for C_{gd1} is slightly more involved. The circuit of Fig. 4.19(*c*) can be redrawn as shown there. In this transformation, the reference or "ground" node was changed. (Which node is called "ground" is arbitrary in an analysis.) Also, the direction and sign of the voltage-controlled current-source were changed. The perceptive reader might notice that the resulting circuit is essentially the same as that used previously to find the output resistance of a source-degenerated current-source. We have

$$v_y = i_3 R_s \tag{4.148}$$

A nodal equation at v_3 gives

$$v_3 G_2 - v_y G_2 - g_{m1} v_y - i_3 = 0 \tag{4.149}$$

Substituting (4.148) into (4.149) and solving for v_3/i_3 gives

$$R_3 \equiv \frac{v_3}{i_3} = R_s(1 + g_{m1}R_2) + R_2 \tag{4.150}$$

Hence, the final time constant is

$$\tau_3 = [R_s(1 + g_{m1}R_2) + R_2]C_{gd1}. \tag{4.151}$$

Taking the –3dB frequency as the inverse of the sum of the time constants results in,

$$\begin{aligned} \omega_{-3dB} &\cong \frac{1}{R_s C_{gs1} + R_2 C_2 + [R_s(1 + g_{m1}R_2) + R_2]C_{gd1}} \\ &= \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2[C_2 + C_{gd1}]} \end{aligned} \tag{4.152}$$

6. This procedure exactly calculates the coefficient of the "s" term in the denominator, b_1 in (4.3). The approximation is in assuming that higher-order terms in the denominator and any zeros can be ignored around the –3dB frequency so that the denominator is approximately $(1 + b_1 s)$ and the –3dB frequency is approximately $1/b_1$.

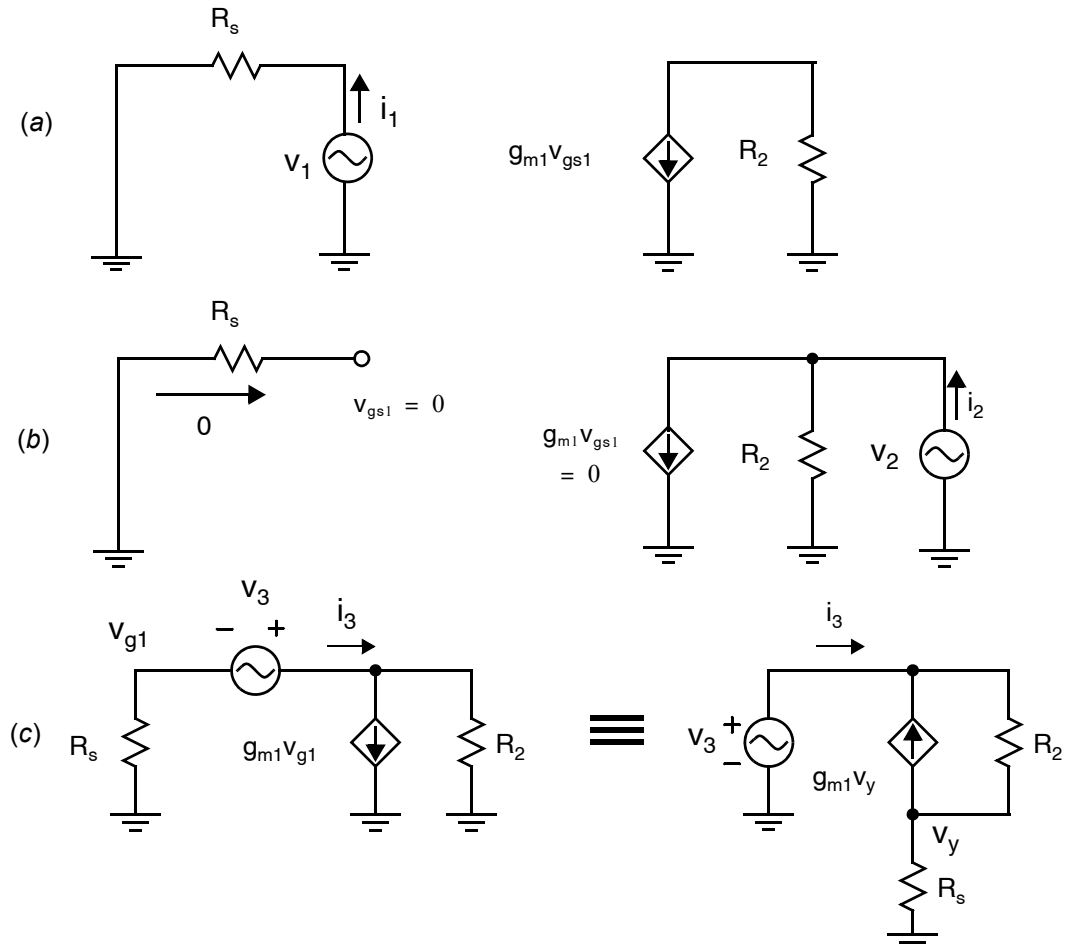


Fig. 4.19 Analysis of a common-source amplifier by the method of zero-value time constants. The complete small-signal schematic appears in Fig. 4.14. (*a*) The simplified small-signal schematic for determining the zero-value time-constant of C_{gs1}. (*b*) Simplified small-signal schematic for determining the zero-value time-constant of C_2. (*c*) Two equivalent small-signal schematics for determining the zero-value time-constant of C_{gd1}.

The bandwidth estimate in (4.152) is exactly the same as (4.131), which was obtained by performing a nodal analysis of the complete circuit and then applying a dominant-pole approximation.[7]

7. This equality is not surprising since (4.131) is predicated on the exact same assumptions that are the basis of the zero-value time-constant method. That is, that the transfer function zeros are negligible and a linear approximation may be made for the denominator: $1 + b_1 s + \ldots + b_n s^n \cong 1 + b_1 s$.

If the load capacitance is very large, the time constant at the output, τ_2, dominates and (4.152) reduces to

$$\omega_{-3dB} \cong \frac{1}{R_2 C_2} \tag{4.153}$$

However, often in analog integrated circuits, the first term in the denominator of (4.152) dominates, yielding the same expression obtained by applying the Miller Theorem and neglecting the output node:

$$\omega_{-3dB} \cong \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + A)]} \tag{4.154}$$

where $A = g_{m1}R_2$ is the magnitude of the low-frequency gain.

EXAMPLE 4.10

Assume all transistors have $W/L = 100\mu m/1.6\mu m$ in Fig. 4.13, and use the transistor parameters listed in Table 1.5 for the 0.8-μm CMOS technology in Table 1.5. Also assume that $I_{bias} = 100\mu A$, $R_s = 180k\Omega$, $C_L = 0.3pF$, $C_{gs1} = 0.2pF$, $C_{gd1} = 0.015pF$, $C_{db1} = 20fF$, $C_{gd2} = 22fF$, and $C_{db2} = 36fF$. Estimate the –3dB frequency of the common-source amplifier of Fig. 4.13.

Solution

We have

$$R_2 = r_{ds1} \| r_{ds2} = 77k\Omega \tag{4.155}$$

and

$$C_2 = C_L + C_{db1} + C_{db2} + C_{gd2} = 0.38pF \tag{4.156}$$

The terms involving R_s, namely $R_s[C_{gs1} + C_{gd1}(1 + A)]$, are equal to 0.26μs. The terms involving R_2, namely $R_2(C_{gd1} + C_2)$ are equal to 0.03μs. The -3dB frequency (in hertz) is approximately.

$$\begin{aligned} f_{-3dB} &\cong \frac{1}{2\pi} \frac{1}{R_{in}[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)} \\ &= 550 \text{ kHz} \end{aligned} \tag{4.157}$$

4.2.5 Common-Source Design Examples

When parasitic MOS capacitances limit the high-frequency response of an amplifier, its bandwidth can be improved by sizing the offending transistor(s) as small as possible. For a fixed current, this implies that the device will exhibit a relatively high value of V_{eff}. However, operation under very high values of V_{eff} is generally avoided since it requires higher dc voltage drops to appear across the transistors to keep them in active mode, and also causes mobility degradation which can decrease amplifier gain.

EXAMPLE 4.11

We are to design the common source amplifier in Fig. 4.13 to provide a gain of 20 while driving a capacitive load of $C_L = 100$ fF with maximal bandwidth. The transistor parameters are those listed in Table 1.5 for the 0.18-μm CMOS technology. The input source resistance is $R_s = 40$ kΩ and the supply voltage is 1.8 V. The ideal current source I_{bias} sinks 50 μA and the total power consumption is to be less than 1 mW.

Solution

SPICE! *Try simulating this design using the provided netlist.*

An approximate expression for the bandwidth ω_{-3dB} is given in (4.152). In this example, the load capacitance is modest and the source resistance high, so the Miller effect is likely to be the major limitation in the amplifier's bandwidth. To maximize ω_{-3dB} we wish to minimize the Miller capacitance $C_{gd1} = C_{ov1}$ which requires us to use a small device width, W_1. As we saw in Chapter 1, this implies relatively large values of $V_{eff,1}$. However, to avoid mobility degradation (which would make it difficult to obtain the targeted gain), we restrict ourselves to a maximum effective gate-source voltage of

$$V_{eff,1} = \frac{1}{2\theta} = \frac{1}{2 \cdot 1.7\ \mathrm{V}^{-1}} \cong 300\ \mathrm{mV}$$

If we ensure that $L_1 \ll L_2$, then $r_{ds2} \gg r_{ds1}$ so that $R_2 \cong r_{ds1}$ in (4.124). Hence, the dc gain $A_0 \cong -g_{m1}r_{ds1}$ is approximately equal to the intrinsic gain of Q_1,

$$A_0 \cong -g_{m1}r_{ds1} = -\frac{2I_{D1}}{V_{eff,1}} \cdot \frac{1}{\lambda I_{D1}} = -\frac{2L_1}{\lambda L_1 V_{eff,1}} \tag{4.158}$$

where we have made a coarse approximation in assuming $r_{ds1} = 1/\lambda I_{D\text{-}sat} \cong 1/\lambda I_D$. Substituting the value $\lambda L_1 = 0.08$ μm/V from Table 1.5 along with $V_{eff,1} = 300$ mV and $|A_0| = 20$ into (4.158) allows us to solve for the device length.

$$L_1 = |A_0|\lambda L_1 V_{eff,1} = (20/2) \cdot 0.08\ \mu\mathrm{m/V} \cdot 0.3\ \mathrm{V} = 0.24\ \mu\mathrm{m}$$

Note that increasing the drain current while maintaining $V_{eff,1} = 300$ mV will increase g_{m1} and reduce r_{ds1} roughly in proportion resulting in approximately the same gain but a smaller R_2 in the denominator of ω_{-3dB} in (4.152). Hence, bandwidth is maximized by using all of the available current. In this case, for a total power consumption of 1 mW and reserving $I_{bias} = 50$ μA of current for the biasing transistor Q_3,

$$I_{D1} = \frac{1\ \mathrm{mW}}{1.8\ \mathrm{V}} - 50\ \mu\mathrm{A} \cong 500\ \mu\mathrm{A}$$

From the desired drain current, gate length, and $V_{eff,1}$ we may compute the required gate width

$$I_{D1} = \frac{1}{2}\mu_n C_{ox}\frac{W_1}{L_1}V_{eff,1}^2$$

$$\Rightarrow W_1 = \frac{2I_{D1}L_1}{\mu_n C_{ox}V_{eff,1}^2} = \frac{2 \cdot 500\ \mu\mathrm{A} \cdot 0.24\ \mu\mathrm{m}}{270\ \mu\mathrm{A/V}^2(0.3\ \mathrm{V})^2} \cong 10\ \mu\mathrm{m}$$

To ensure $L_2 \gg L_1$, we take $L_2 = 3L_1 = 0.72\ \mu m$. Since they have the same drain currents, the width of Q_2 may be conveniently taken 3 times that of Q_1, although this is not critical.

$$W_2 = 3 \cdot 10\ \mu m = 30\ \mu m$$

Finally, Q_3 is sized to provide the desired current ratio in the current mirror formed with Q_2.

$$L_3 = L_2 = 0.72\ \mu m$$

$$W_3 = W_2 \cdot \left(\frac{50\ \mu A}{500\ \mu A}\right) = 3\ \mu m$$

When the Miller effect is dominant, as in Example 4.11, the amplifier bandwidth is maximized by taking a small device width and high V_{eff} for common-source transistor Q_1. However, large device widths and small values of V_{eff} are preferable when a large load capacitance causes the output time constant to dominate.

EXAMPLE 4.12

We wish to design the common source amplifier in Fig. 4.13 with minimal power consumption while providing a 3-dB bandwidth of 5 MHz and a gain of at least 20 using the transistor parameters listed in Table 1.5 for the 0.18-μm CMOS technology with $C_L = 10$ pF and $R_s = 1\ k\Omega$. The ideal current source I_{bias} is 50 μA.

Solution

In this case, the load capacitance is large and the source resistance lower than in Example 4.11. The output time constant will dominate.

$$\omega_{-3dB} \cong 1/(r_{ds,1} || r_{ds,2})C_L = 2\pi \cdot 5 \cdot 10^6 \text{ rad/sec} \qquad (4.159)$$

Substituting the fixed $C_L = 10$ pF into (4.159) yields,

$$(r_{ds,1} || r_{ds,2}) = 3.2\ k\Omega$$

Note that $(r_{ds,1} || r_{ds,2}) = 1/I_D(\lambda_1 + \lambda_2)$. To minimize I_D and, hence, power consumption we require large λ_1 and λ_2 which in turn demands small L_1 and L_2. Therefore, we use the minimum possible in this CMOS process.

$$L_1 = L_2 = 0.18\ \mu m$$

Using the values for λL from Table 1.5,

$$\lambda_1 = \lambda_2 = (0.08\ \mu m/V)/0.18\ \mu m = 0.44\ V^{-1}$$

From this, the required drain current is found.

$$I_D = \frac{1}{r_{ds} \cdot \lambda} \cong 350\ \mu A$$

The result has been rounded to the nearest integer multiple of the bias current, 50 μA, in order to simplify the current mirror design. To meet the gain requirement,

$$|A_0| \cong g_{m1}(r_{ds,1} || r_{ds,2}) = \frac{2I_D}{V_{eff,1}} \cdot \frac{1}{(\lambda_1 + \lambda_2)I_D} = \frac{2}{V_{eff,1}} \cdot \frac{1}{(\lambda_1 + \lambda_2)} > 20$$

$$\Rightarrow V_{eff,1} < \frac{2}{20(0.44\ V^{-1} + 0.44\ V^{-1})} = 114\ mV$$

For some margin, we take $V_{eff,1} = 100$ mV and the resulting transistor width is

$$W_1 = \frac{2I_{D1}L_1}{\mu_n C_{ox} V_{eff,1}^2} = \frac{2 \cdot 355\ \mu A \cdot 0.18\ \mu m}{270\ \mu A/V^2 (0.1\ V)^2} \cong 47 \mu m$$

The width of Q_2 may be chosen the same, $W_2 = 47\ \mu m$ and the size of Q_3 chosen to provide the correct current ratio in the current mirror formed by Q_2 and Q_3:

$$\frac{(W_3/L_3)}{(W_2/L_2)} = \frac{50\ \mu A}{350\ \mu A}$$

$$\Rightarrow L_3 = 0.18\ \mu m \quad \text{and} \quad W_3 = 6.7\ \mu m$$

Note that the gain of the amplifier in Example 4.12 is maximized by increasing W_1 to operate Q_1 near subthreshold.

4.2.6 Common-Gate Amplifier

The frequency response of the common-gate stage is usually superior to that of the common-source stage for two reasons. First, there is no Miller Capacitance coupling from the input node to the high-gain output node. Second, assuming R_L is not considerably larger than r_{ds}, the common-gate stage exhibits a low input resistance at the source node. The small-signal circuit of a common-gate amplifier being driven by a small signal Norton-equivalent input current source, I_{in}, is shown in Fig. 4.20. All parasitic capacitances between the drain and small-signal ground have been combined into C_2 and the body-effect may be included by an additional transconductane g_s in parallel with g_m. We will use the technique of zero-value time-constant analysis to estimate the bandwidth of this amplifier.

First we will estimate the time-constant associated with C_{gs} by analysis of the circuit in Fig. 4.21 where the input source and all other capacitances have been set equal to zero. The capacitor C_{gs} is connected between the source and small-signal ground and may, therefore, include C_{sb}. The small-signal resistance seen by it is simply the resistance seen at the source, $R_1 = r_{in} || R_s$. The input resistance of the common-gate amplifier was previously shown to be

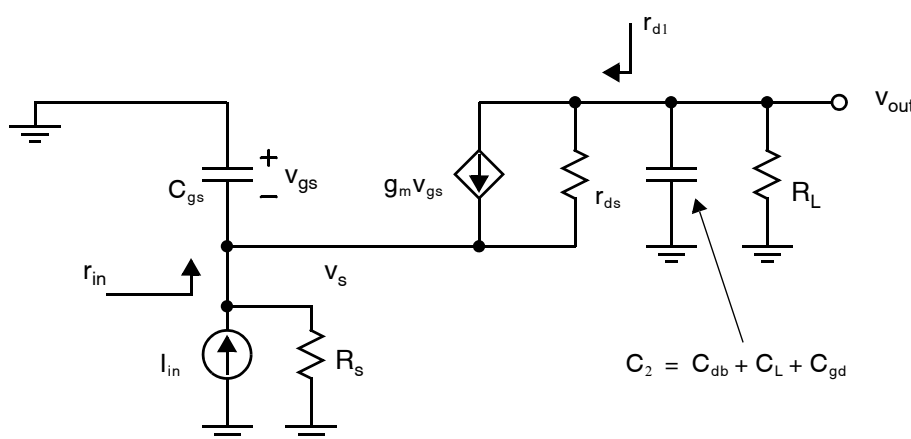


Fig. 4.20 The high-frequency small-signal model for a common-gate amplifier.

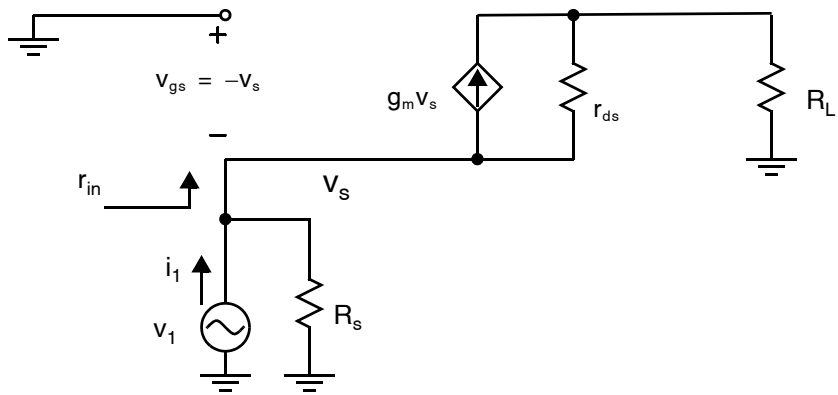


Fig. 4.21 A simplified small-signal model for computing the time-constant of C_{gs}.

$$r_{in} \cong \frac{1}{g_m}\left(1 + \frac{R_L}{r_{ds}}\right) \tag{4.160}$$

Under the assumption that R_L is not too much bigger than r_{ds}, we have

$$R_1 = r_{in} \| R_s \cong \frac{1}{g_m} \| R_s \cong \frac{1}{g_m} \tag{4.161}$$

Thus, the time-constant associated with C_{gs} is

$$\tau_1 = (r_{in} \| R_s) C_{gs} \cong \frac{C_{gs}}{g_m} \tag{4.162}$$

This assumes $r_{in} \ll R_s$, but in the event this assumption is not justified, the time constant would be smaller and therefore less important.

Consider next the resistance seen-by C_2 with all other capacitances set to zero. Once again this capacitor is grounded. The resistance seen by it is

$$R_2 = R_L \| r_{d1} \tag{4.163}$$

where r_{d1} is the resistance looking into the drain with source-degeneration resistor R_S in place. From our previous analysis of the output impedance of a source-degenerated current source at low frequencies, we know this resistance is given by

$$r_{d1} = r_{ds}(1 + g_m R_s) \tag{4.164}$$

Often, this impedance will be much greater than R_L in which case $R_2 \cong R_L$ and

$$\tau_2 = (R_L \| r_{d1}) C_2 \cong R_L C_2 \tag{4.165}$$

The -3dB frequency will be approximately given by $1/(\tau_1 + \tau_2)$. Note the absence of any time constants having capacitive terms effectively multiplied by the low-frequency gain of the amplifier unlike the common-source stage. Whereas in the common-source amplifier C_{gd} appears between two nodes related by a large negative gain, in the common-gate amplifier C_{gd} is grounded so there is no Miller effect. This results in significantly superior high-frequency operation.

In the next section we will see that by combining a common-gate stage with a common-source stage, even further improvements are achieved. This combination mitigates the Miller effect, like the common-gate stage, but has the high input impedance of the common-source stage.

Key Point: *All small-signal capacitances in the common-gate amplifier are (small-signal) grounded. Hence there is no Miller effect and it can provide high-frequency performance superior to that of the common-source amplifier, although with much lower input resistance.*

4.3 CASCODE GAIN STAGE

We next consider the frequency response of the cascode amplifiers shown in Fig. 4.22. Comparing the telescopic cascode of Fig. 4.22(*a*) and the folded cascode of Fig. 4.22(*b*), we see that the folded-cascode amplifier suffers from smaller carrier mobility in the common gate transistor Q_2. This results in a lower transconductance, g_{m2}, a larger time-constant associated with C_{gs2}, as seen in (4.162), and generally lower bandwidth.

Recall that a major reason for the popularity of cascode stages is that they can have quite large gain for a single stage due to the large impedances at the output. Normally this high gain is obtained without any degradation in speed, or sometimes with an improvement in speed. If, on the other hand, the current source at the output node has a modest or low output resistance (typical when they are realized with just a simple resistor) the cascode stage can provide improved bandwidth without any degradation in gain, or sometimes a small improvement in gain.

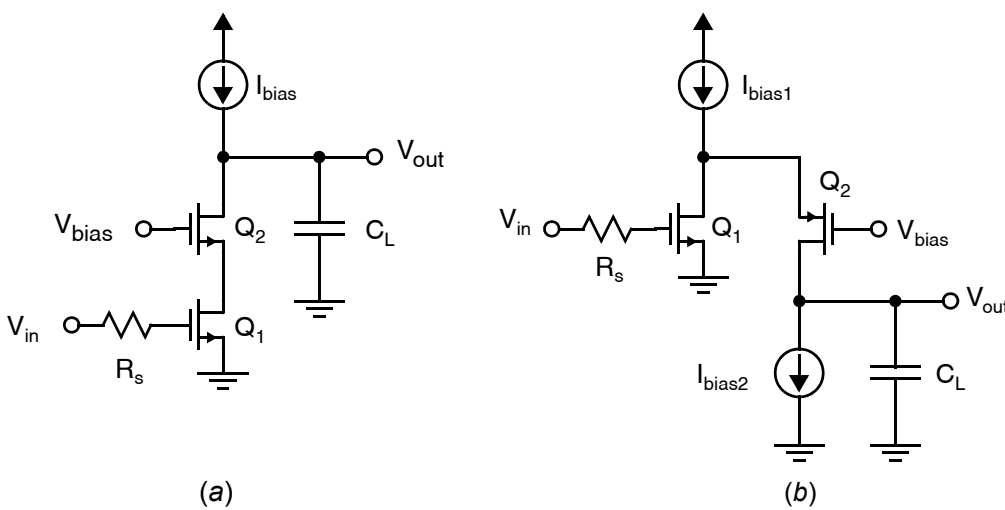


Fig. 4.22 (*a*) A telescopic cascode amplifier and (*b*) a folded-cascode amplifier.

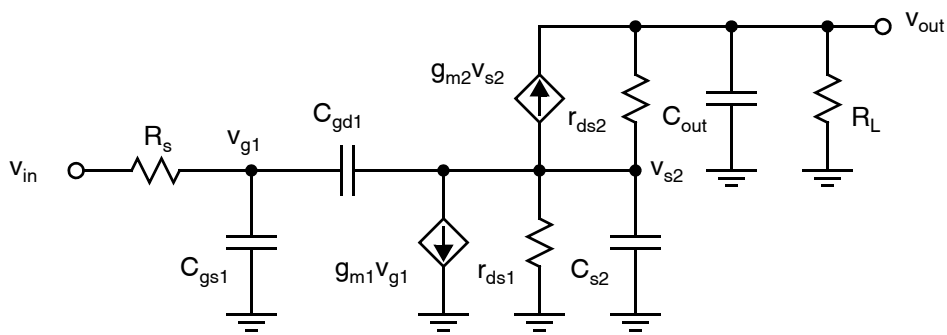


Fig. 4.23 The small-signal model of the cascode gain stage.

Key Point: *The cascode gain stage uses a common-gate transistor Q_2 to reduce the V_{DS} variations on a common-source transistor Q_1. The result is high output resistance providing potentially high gain, a reduced Miller effect, and reduced short-channel effects. However, the available output voltage swing is reduced.*

The exact high-frequency analysis of a cascode gain stage is usually left to simulation on a computer, however an approximate analysis based on zero-value time-constants is not too complicated. As before, in the zero-value time-constant analysis all independent sources are set to zero (i.e. here, v_{in} is set to 0 volts) and each capacitor is considered in turn with all other capacitors set to zero. The corresponding time-constants are found and labeled τ_k. Then the –3dB frequency, ω_{-3dB}, is estimated to be one over the sum of all the time-constants. This technique gives some insight into the relative importance of each capacitor in determining the overall –3dB frequency.

The small-signal model being analyzed is shown in Fig. 4.23 where the total capacitance at the output node, C_{out}, is the parallel combination of $C_{gd2} + C_{db2}$, the load capacitance C_L, and the output capacitance of the bias current source, C_{bias}.

$$C_{out} = C_{gd2} + C_{db2} + C_L + C_{bias} \tag{4.166}$$

All capacitances at the source of Q_2 are combined into

$$C_{s2} = C_{db1} + C_{sb2} + C_{gs2} \tag{4.167}$$

The resistance seen by C_{out} is the amplifier's output resistance, R_{out}. Hence, the output time-constant is

$$\tau_{out} = R_{out}C_{out} = (r_{d2} \| R_L)C_{out} \tag{4.168}$$

The resistance seen by C_{gs1} is R_s and the corresponding time-constant is simply

$$\tau_{gs1} = C_{gs1}R_s \tag{4.169}$$

The resistance seen by C_{s2} is the parallel combination of r_{in2} defined in (3.51) and r_{ds1}.

$$g_{s2} = 1/r_{s2} = g_{in2} + g_{ds1} \tag{4.170}$$

Hence, its time constant is

$$\tau_{s2} = (r_{in2} \| r_{ds1})C_{s2} = \frac{C_{s2}}{g_{in2} + g_{ds1}} \tag{4.171}$$

The calculation of the time-constant corresponding to C_{gd1} is more involved, but identical to that performed for C_{gd1} in the common-source amplifier in Example 4.10. The result is given by (4.151) where R_2 is the total resistance at the drain of Q_1, in this case given by $r_{in2}\|r_{ds1}$.

$$\begin{aligned}\tau_{gd1} &= \{R_s[1 + g_{m1}(r_{in2}\|r_{ds1})] + (r_{in2}\|r_{ds1})\}C_{gd1} \\ &\cong R_s[1 + g_{m1}(r_{in2}\|r_{ds1})]C_{gd1}\end{aligned} \tag{4.172}$$

The approximation in (4.172) assumes $g_{m1}R_s \gg 1$. Since the low-frequency gain from v_{g1} to v_{s2} is $-g_{m1}(r_{in2}\|r_{ds1})$, you may recognize this time constant as a manifestation of the Miller effect. Notice that C_{gd1} is multiplied by one minus the low-frequency gain between the two nodes it bridges, as illustrated in Fig. 4.24. If R_s is large, say on the order of a transistor output impedance r_{ds}, and I_{bias} has an output impedance on the order of $g_m r_{ds}^2$, then this time-constant is approximately given by

$$\tau_{gd1} \approx C_{gd1}\frac{g_m r_{ds}^2}{2} \tag{4.173}$$

which is almost as large as the corresponding time-constant for a common-source amplifier—a fact not well known.

The sum of the time-constants is then given by

$$\begin{aligned}\tau_{total} &= \tau_{out} + \tau_{gs1} + \tau_{s2} + \tau_{gd1} \\ &\approx (g_{m2}r_{ds1}r_{ds2}\|R_L)C_{out} + C_{gs1}R_s + \\ &\quad +(r_{in2}\|r_{ds1})C_{s2} + R_s[1 + g_{m1}(r_{in2}\|r_{ds1})]C_{gd1}\end{aligned} \tag{4.174}$$

The -3 dB frequency is estimated to be one over this time-constant (i.e. $\omega_{-3dB} = 1/\tau_{total}$).

EXAMPLE 4.13

Estimate the –3dB frequency of the cascode amplifier of Fig. 4.22(*a*). Assume that the current source I_{bias} has a high output impedance, on the order of $R_L \approx g_{m\text{-}p}r_{ds\text{-}p}^2$. Further assume that for all transistors, g_m = 1 mA/V,

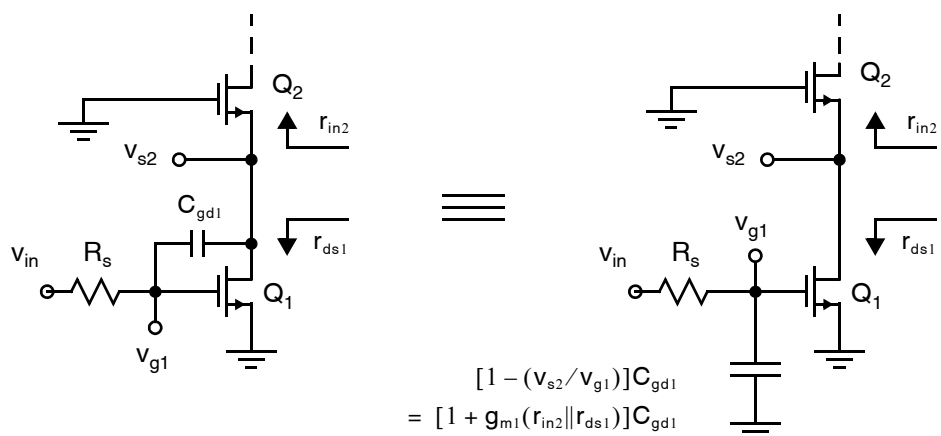


Fig. 4.24 The Miller effect for the calculation of the zero-value time-constant τ_{gd1} in a cascode amplifier.

$r_{ds} = 100\ k\Omega$, $C_{gs} = 0.2$ pF, $C_{gd} = 15$ fF, $C_{sb} = 40$ fF, and $C_{db} = 20$ fF. The other component values are $R_s = 180\ k\Omega$, $C_L = 5$ pF, and $C_{bias} = 20$ fF.

Solution

The time-constants associated with each capacitor are evaluated and summed to form the bandwidth estimate. First note that

$$\begin{aligned} C_{s2} &= C_{db1} + C_{sb2} + C_{gs2} = 0.26\ \text{pF} \\ C_{out} &= C_{gd2} + C_{db2} + C_L + C_{bias} = 5.055\ \text{pF} \end{aligned} \tag{4.175}$$

It was shown in Section 3.7 that when I_{bias} has high output resistance, $r_{in2} \approx r_{ds}$. Hence, we have

$$\begin{aligned} \tau_{gs1} &= R_s C_{gs1} = 36\ \text{ns} \\ \tau_{gd1} &\approx \frac{g_m r_{ds}^2}{2} C_{gd1} = 75\ \text{ns} \\ \tau_{s2} &\approx \frac{r_{ds}}{2} C_{s2} = 13\ \text{ns} \\ \tau_{out} &\approx (g_m r_{ds}^2 \| g_m r_{ds}^2) C_{out} = 25.3\ \mu\text{s} \end{aligned} \tag{4.176}$$

An estimate of the –3dB bandwidth is obtained by taking the inverse of the sum of all time constants in (4.176),

$$\omega_{-3dB} \cong \frac{1}{\tau_{gs1} + \tau_{gd1} + \tau_{s2} + \tau_{out}} = 2\pi \times 6.3\ \text{kHz} \tag{4.177}$$

When used to provide a large low-frequency gain, the cascode amplifier must be biased by a high-quality current source, I_{bias} whose output resistance is on the order of $R_L \approx g_{m\text{-}p} r_{ds\text{-}p}^2$. In that case, the overall gain is

$$A_V \approx -\frac{1}{2} g_m^2 r_{ds}^2 = -\frac{1}{2}\left(\frac{g_m}{g_{ds}}\right)^2 \tag{4.178}$$

and the output resistance is

$$R_{out} = R_L \| r_{d2} \approx \frac{1}{2} g_m r_{ds}^2 \tag{4.179}$$

When designed for large output resistance like this, and especially when there is a large load capacitance C_L, the output time constant τ_{out} expressed in (4.168) will generally dominate over all others. We can substitute $R_{out} \approx g_m r_{ds}^2/2$ into (4.168) to obtain a rough estimate of this time constant,

$$\tau_{out} \approx g_m r_{ds}^2 C_{out}/2 \tag{4.180}$$

and the –3dB frequency is approximately equal to its inverse,

$$\omega_{-3dB} \cong \frac{1}{\tau_{out}} = \frac{1}{R_{out} C_{out}} \approx \frac{2 g_{ds}^2}{g_m C_{out}} \tag{4.181}$$

EXAMPLE 4.14

Reconsider Example 4.13, but this time neglect all time constants other than the output time constant.

Solution

Using (4.181) to obtain a bandwidth estimate yields an almost identical result to that obtained in (4.177) considering all time-constants.

$$\omega_{-3dB} \cong 1/\tau_{out} = 2\pi \times 6.3 \text{ kHz} \tag{4.182}$$

This is not surprising since the large load capacitance $C_L = 5pF$ ensures that the time-constant at the output node dominates. Looking at (4.176), the second most important time-constant is τ_{gd1} whose value is almost 3 orders of magnitude smaller than τ_{out}. Hence, its effect on the -3dB frequency is negligible.

When one pole dominates, as in Example 4.14, we can reasonably model the amplifier frequency response over a wide frequency range using a dominant pole approximation,

$$A(s) \cong \frac{A_v}{1 + (s/\omega_{-3dB})} \tag{4.183}$$

When the cascode amplifier is part of a feedback loop, we shall see that the frequency band of operation is primarily at frequencies substantially larger than ω_{-3dB}. In this range, the (s/ω_{-3dB}) term will dominate in the denominator and using (4.178) and (4.181), the frequency response is approximated by

$$A(s) \cong \frac{A_v}{s/\omega_{-3dB}} \cong -\frac{g_{m1}}{sC_L} \tag{4.184}$$

The approximations of (4.181) and (4.184) are quite good unless either the source impedance or source capacitance is very large.

The dominant pole estimated in (4.181) is due to the parallel combination of the load capacitance, C_L, and large output resistance, $R_{out} \approx (g_m r_{ds}^2)/2$. A more accurate analysis would reveal two other poles:

- ω_{in} due to the parallel combination of R_s and total capacitance at v_{g1}, which includes a Miller capacitance
- ω_{s2} due to the parallel combination of C_{s2} and the impedance at v_{s2}

Both depend, in no small measure, on the admittance looking up into the source of of Q_2, Y_{in2}, depicted in Fig. 4.25. This admittance can be found by making use of (3.51) where R_L is replaced by $R_L \| (1/sC_L)$. Such a substitution results in

$$\begin{aligned} Y_{in2} &= \frac{g_{m2} + g_{s2} + g_{ds2}}{1 + \dfrac{g_{ds2}R_L}{1 + sR_LC_L}} \\ &\cong g_{m2}\left(\frac{1 + sR_LC_L}{1 + sR_LC_L + g_{ds2}R_L}\right) \\ &\cong g_{m2}\left(\frac{1 + sR_LC_L}{g_{ds2}R_L + sR_LC_L}\right) \end{aligned} \tag{4.185}$$

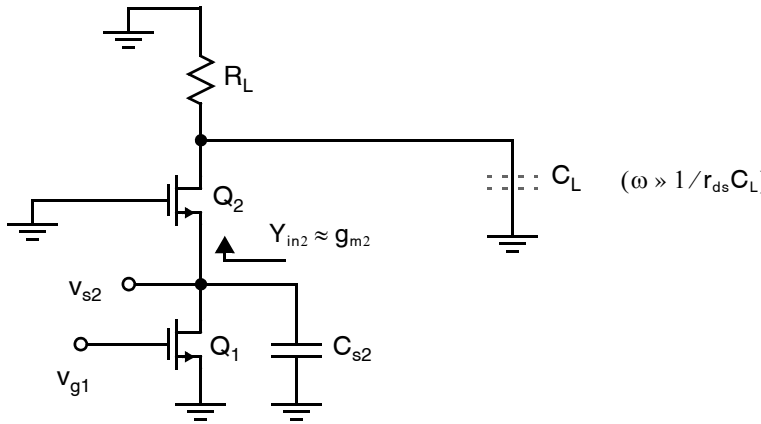


Fig. 4.25 A large load capacitance leads to a reduced impedance (increased admittance) looking into the source of Q_2. This, in turn, ameliorates the Miller effect at the input, and increases the frequency of the pole due to C_{s2}.

At frequencies where $\omega \gg 1/(r_{ds}C_L)$, the terms in s dominate and $Y_{in2} \cong g_{m2}$. Intuitively, this is because at such frequencies the load capacitor effectively shorts the output to ground, and we therefore see only a small-signal resistance $1/g_{m2}$ looking into the source of Q_2, as illustrated in Fig. 4.25.

Since the high-frequency impedance at the node v_{s2} has been reduced to $1/g_{m2}$, the gain from v_{g1} to v_{s2} at these frequencies is roughly just $-g_m/g_m = -1$. Hence, the effective Miller capacitance at the input is only $2C_{gd1}$, resulting in an input pole of

$$\omega_{in} \cong \frac{1}{R_s(C_{gs1} + 2C_{gd1})} \tag{4.186}$$

The approximate frequency of the pole due to the parallel combination of C_{s2} and the impedance at v_{s2} is then simply given by

$$\omega_{s2} \cong \frac{g_{m2}}{C_{s2}} \tag{4.187}$$

When the source resistance, R_S, is small (for example, when the cascode stage is driven by an opamp circuit), $\omega_{in} \gg \omega_{s2}$ and the second pole of a cascode amplifier is approximately given by (4.187).

It is very easy to derive a upper bound on ω_{s2}. From (4.167) we see that

$$C_{s2} > C_{gs2} = \frac{2}{3}(WL)C_{ox} \tag{4.188}$$

For a folded-cascode amplifier with PMOS Q_2,

$$g_{m2} = \mu_p C_{ox}\left(\frac{W}{L}\right)_2 V_{eff2} \tag{4.189}$$

For the NMOS telescopic-cascode amplifier replace μ_n for μ_p in (4.189). Substituting (4.188) and (4.189) into (4.187) gives the approximate frequency of the second pole.

$$\omega_{s2} < \frac{\mu_p C_{ox}(W/L)_2 V_{eff2}}{(2/3)(WL)_2 C_{ox}} = \frac{3\mu_p V_{eff2}}{2L_2^2} \tag{4.190}$$

When enclosed in a feedback loop, we shall see that stability demands the unity gain frequency of an amplifier be restricted to less than the frequency of its second pole. Equation (4.190) shows that for cascode amplifiers with a dominant output pole, the second pole frequency is relatively independent of that actual design once V_{eff2} is chosen which is usually determined by maximum voltage swing requirements. Hence, this equation is an upper-limit on the unity-gain frequency of any feedback amplifier that uses a cascode gain stage**.** Note the very strong dependance on the channel length, L_2. Finally, you may recognize (4.190) as nothing more than the intrinsic speed (unity-gain frequency) of the cascode-transistor, Q_2.

> ***Key Point:*** *The unity-gain frequency of any feedback amplifier that uses a cascode gain stage with a dominant output pole is limited by the unity-gain frequency of the cascode transistor,* Q_2

$$\omega_{s2} < 2\pi f_{T,2} \tag{4.191}$$

EXAMPLE 4.15

Estimate the upper-bound on the frequency of the second-pole of a folded-cascode amplifier with a large load capacitance for a $0.18\mu m$ technology where a typical value of $0.25V$ is chosen for V_{eff2}.

Solution

Normally, a minimum length of a cascode transistor in an analog circuit might be 25% to 50% larger than the minimum length of transistors used in digital circuits. Therefore, assuming $L_2 = 1.5 \cdot 0.18\mu m = 0.27\mu m$, and using $\mu_p = 0.0082\ m^2/Vs$ and $V_{eff2} = 0.25V$ in (4.190) gives

$$\omega_{p2} < 42.2\times10^9 \text{rad} = 2\pi \cdot 6.7\text{GHz}$$

In most practical opamp designs, the unity-gain frequency of a typical design might be limited to around one-quarter the frequency of the upper-bound established by (4.190) due to the uncertainty and temperature dependence of the variables there, or in this case around 1.5 GHz. For a NMOS telescopic-cascode amplifier, the upper-bound would be 2-4 times higher due to the higher mobility of electrons compared to holes.

4.4 SOURCE-FOLLOWER AMPLIFIER

The high-frequency analysis of source-follower amplifiers is somewhat involved. It will be shown that these types of amplifiers can have complex poles and thus a designer should be careful that the circuit does not exhibit too much overshoot and ringing. Also shown is a compensation circuit that will result in only real axis poles and therefore no overshoot and ringing.

We shall find the frequency response of the source follower for the common situation in integrated circuits where the source may be modeled by its Norton equivalent circuit and the load is purely capacitive as shown in Fig. 4.26. The small-signal model for this circuit, including the parasitic capacitances, is shown in Fig. 4.27.

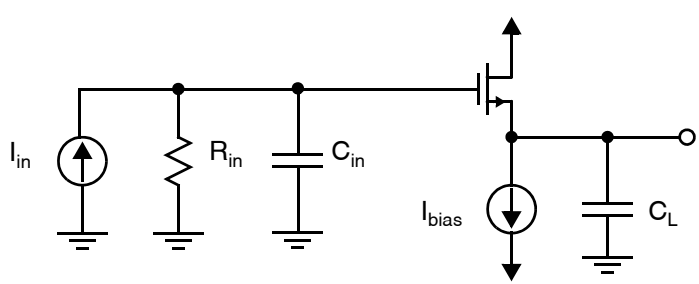


Fig. 4.26 The configuration used to analyze the frequency response of the source follower.

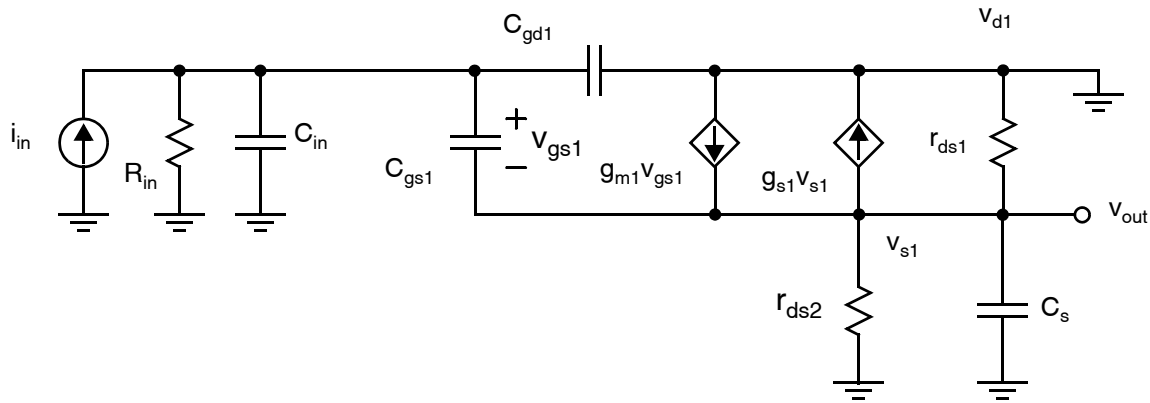


Fig. 4.27 An equivalent small-signal model for the source-follower.

Capacitor C_s includes both the load capacitor, C_L, and the parasitic capacitor C_{sb1}. Similar to what was done at low-frequencies, r_{ds1}, r_{ds2}, and the voltage-controlled current source modelling the body-effect current-source can be modelled by a single resistor. This allows us to analyze the simplified small-signal model shown in Fig. 4.28 where again $R_{s1} = r_{ds1} \| r_{ds2} \| (1/g_{s1})$ and the input capacitance is given by $C'_{in} = C_{in} + C_{gd1}$.

Nodal analysis is possible, but it is very complicated for this example. The analysis will proceed in three steps. First, the gain from v_{g1} to v_{out} will be found. Next, the admittance, Y_g, looking into the gate of Q_1, but not taking into account C_{gd1}, will be found. This will be used to find the gain from i_{in} to v_{g1}. Finally, the overall gain from v_{in} to v_{out} will be found and the results interpreted.

The nodal equation at v_{out} is

$$v_{out}(sC_s + sC_{gs1} + G_{s1}) - v_{g1}sC_{gs1} - g_{m1}(v_{g1} - v_{out}) = 0 \tag{4.192}$$

Solving for v_{out}/v_{g1}, we have

$$\frac{v_{out}}{v_{g1}} = \frac{sC_{gs1} + g_{m1}}{s(C_{gs1} + C_s) + g_{m1} + G_{s1}} \tag{4.193}$$

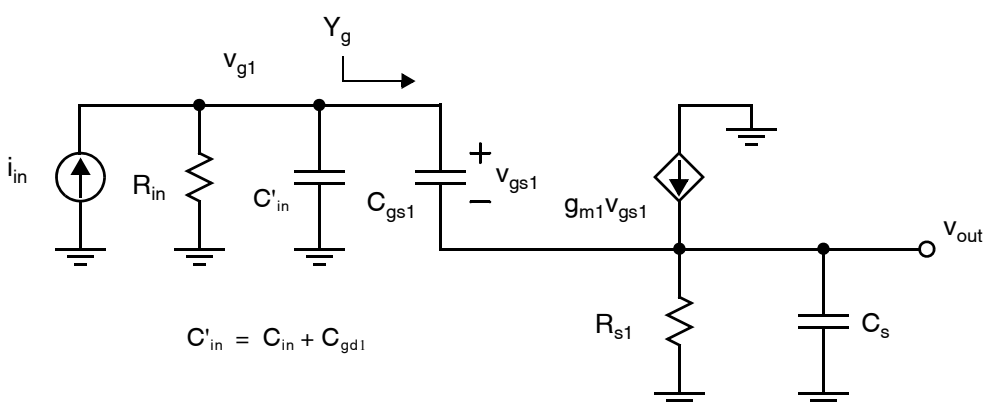


Fig. 4.28 A simplified equivalent small-signal model for the source-follower.

The next step is to calculate the admittance, Y_g, looking into the gate of Q_1, but not taking into account the current going into C_{gd1}. The input current is given by

$$i_{g1} = (v_{g1} - v_{out})sC_{gs1} \tag{4.194}$$

Using (4.193) to eliminate v_{out} in (4.194) and solving for $Y_g = i_{g1}/v_{g1}$, we have

$$Y_g = \frac{i_{g1}}{v_{g1}} = \frac{sC_{gs1}(sC_s + G_{s1})}{s(C_{gs1} + C_s) + g_{m1} + G_{s1}} \tag{4.195}$$

Also, an equation can be written relating the input current, i_{in}, to the gate voltage, v_{g1}, as

$$i_{in} = v_{g1}(sC'_{in} + G_{in} + Y_g) \tag{4.196}$$

Substituting (4.195) into (4.196) and rearranging gives

$$\frac{v_{g1}}{i_{in}} = \frac{s(C_{gs1} + C_s) + g_{m1} + G_{s1}}{a + sb + s^2c} \tag{4.197}$$

where

$$\begin{aligned} a &= G_{in}(g_{m1} + G_{s1}) \\ b &= G_{in}(C_{gs1} + C_s) + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1} \\ c &= C_{gs1}C_s + C'_{in}(C_{gs1} + C_s) \end{aligned} \tag{4.198}$$

Using (4.193) and (4.197), we then have

$$A(s) = \frac{v_{out}}{i_{in}} = \frac{sC_{gs1} + g_{m1}}{a + sb + s^2c} \tag{4.199}$$

Thus, we see that the transfer-function is second-order. Specifically, it has two poles (roots of the denominator) which may be either real or complex-conjugate. If they are complex-conjugate, then the step response of the circuit will exhibit over-shoot and possibly ringing, as described in Section 4.1.5. This potential problem is a disadvantage when using source-followers.

To determine if the transfer function will exhibit ringing, (4.199) can be written in the form

$$A(s) = A(0)\frac{N(s)}{1 + \frac{s}{\omega_0 Q} + \frac{s^2}{\omega_0^2}} \tag{4.200}$$

where ω_0 and Q can be found by equating the coefficients of (4.200) to the coefficients of (4.199). As in Section 4.1.3, parameter ω_o is the resonant frequency and parameter Q is the Q-factor. It is well known that if $Q < \sqrt{1/2} \approx 0.707$, then the magnitude of the transfer-function will have its maximum at dc and there will be no peaking (assuming the zero is at a very-high frequency and therefore has negligible effect). Furthermore, for $Q = \sqrt{1/2}$, the –3dB frequency is equal to ω_o. When the time-domain response is investigated, restrictions on the Q-factor can also be found to guarantee no overshoot for a step input. Specifically, for there to be no overshoot in the step-response, it is necessary that both poles be real which is equivalent to the requirement that $Q \le 0.5$. In the case where $Q > 0.5$, the percentage overshoot of the output voltage can be shown to be given by

$$\%\ \text{overshoot} = 100e^{-\frac{\pi}{\sqrt{4Q^2 - 1}}} \tag{4.201}$$

For the source-follower, equating the coefficients of (4.200) to the coefficients of (4.199) and solving for ω_0 and Q results in

$$\omega_0 = \sqrt{\frac{G_{in}(g_{m1} + G_{s1})}{C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)}} \tag{4.202}$$

$$Q = \frac{\sqrt{G_{in}(g_{m1} + G_{s1})[C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)]}}{G_{in}C_s + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1}} \tag{4.203}$$

If Q is less than 0.5, the poles will be real and distinct. Although this Q equation is rather complex, it is interesting to note that if C_s and/or C'_{in} becomes large (i.e. a large load and/or input capacitor), then Q becomes small and there will be no overshoot (though the circuit will be slow). For example, one possibility is that the load is purely capacitive and very large so that $G_{s1} \ll g_{m1}$ and $C_s/g_{m1} \gg R_{in}(C_{gs1} + C_{in}')$. In this case, it may be shown that $Q \ll 0.5$ and, from (4.72), we have a dominant pole at

$$\omega_{p1} \cong Q\omega_0 = \frac{G_{in}(g_{m1} + G_{s1})}{G_{in}C_s + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1}} \cong \frac{g_{m1}}{C_s} \tag{4.204}$$

Hence, the bandwidth is determined mainly by the zero-value time constant corresponding to the load capacitance C_s.

If Q is greater than 0.5, the poles will be complex-conjugate and the circuit will exhibit overshoot. For example, when C'_{in} and G_{s1} become small[8] then the circuit will have a large Q (i.e., large ringing) when G_{in} becomes small and $C_s \approx C_{gs1}$. Fortunately, the parasitic capacitances and output impedances in practical microcircuits typically result in only moderate overshoot for worst-case conditions.

Finally, note also that the numerator zero of the transfer-function is on the negative real-axis at a frequency given by

$$-\omega_z = -\frac{g_{m1}}{C_{gs1}} \tag{4.205}$$

and is typically at a much higher frequency than ω_o.

EXAMPLE 4.16

Using the parameters in Example 3.4, and also assuming $R_{in} = 18\ k\Omega$, $C_L = 0.2$ pF, and $C_{in} = 10$ fF, find ω_0, Q and the frequency of the zero for the source-follower of Fig. 4.26.

Solution

From Example 3.4 we know $g_{m1} = 0.735$ mA/V, $r_{ds1} = 25\ k\Omega$, $r_{ds2} = 25\ k\Omega$, and $g_{s1} = 0.1$ mA/V. Moreover, we can estimate the parasitics as follows:

$$\begin{aligned} C_{gs1} &= \frac{2}{3}WLC_{ox} + WL_{ov}C_{ox} \\ &= \frac{2}{3}(2\ \mu m)(0.2\ \mu m)(8.5\ fF/\mu m^2) + (2\ \mu m)(0.35\ fF/\mu m) \\ &= 3.0\ fF \end{aligned} \tag{4.206}$$

8. G_{s1} becomes small when the transistor's source is connected to its body terminal which eliminates the body effect.

$$C_{gd1} = WL_{ov}C_{ox} = (2\ \mu m)(0.35\ fF/\mu m) = 0.7\ fF \tag{4.207}$$

$$C_{sb1} = (2\ \mu m)(0.5\ fF/\mu m) = 1\ fF \tag{4.208}$$

Thus, we have

$$C'_{in} = C_{in} + C_{gd1} = 10.7\ fF \tag{4.209}$$

$$G_{s1} = g_{s1} + g_{ds1} + g_{ds2} = 0.18\ mA/V \tag{4.210}$$

$$C_s = C_L + C_{sb1} \cong 0.2\ pF \tag{4.211}$$

and so we can find ω_o as

$$\omega_0 = \sqrt{\frac{G_{in}(g_{m1} + G_{s1})}{C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)}}$$

$$= (4.27\times10^9 rad/s = 2\pi \times 680\ MHz) \tag{4.212}$$

$$Q = \frac{\sqrt{G_{in}(g_{m1} + G_{s1})[C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)]}}{G_{in}C_s + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1}}$$

$$= 0.554 \tag{4.213}$$

This results in an overshoot for a step input given by

$$\%\ \text{overshoot} = 100e^{-\frac{\pi}{\sqrt{4Q^2-1}}} = 0.13\% \tag{4.214}$$

The zero frequency is found using (4.205) to be 39 GHz and thus it can almost certainly be ignored.

When complex conjugate poles occur, they can be eliminated by adding a compensation network. To see this, note that (4.195) can be rewritten as

$$Y_g = sC_2 + \frac{1}{-R_1 - \dfrac{1}{sC_1}} \tag{4.215}$$

where

$$\begin{aligned} C_1 &= \frac{C_{gs1}(C_s g_{m1} - C_{gs1}G_{s1})}{(g_{m1} + G_{s1})(C_{gs1} + C_s)} \cong \frac{g_{m1}C_{gs1}C_s}{(g_{m1} + G_{s1})(C_{gs1} + C_s)} \\ R_1 &= \frac{(C_{gs1} + C_s)^2}{C_{gs1}(C_s g_{m1} - C_{gs1}G_{s1})} \cong \frac{(C_{gs1} + C_s)^2}{C_{gs1}C_s g_{m1}} \\ C_2 &= \frac{C_{gs1}C_s}{C_{gs1} + C_s} \end{aligned} \tag{4.216}$$

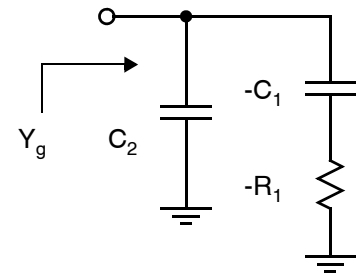


Fig. 4.29 A circuit having the same admittance as the input impedance looking into the gate of a source-follower (ignoring C_{gd}

Key Point: *Source follower circuits can exhibit large amounts of overshoot and ringing under certain conditions. When designing source-followers, the recommended procedure is to check to see if the poles are complex using either (4.203) or a SPICE transient analysis to look for overshoot. When the poles are complex, then increase the load and/or input capacitor or, alternatively, add the compensating network shown in Fig. 4.30.*

and the approximation is due to the fact that typically $C_s > C_{gs1}$ and $g_{m1} > G_{s1}$. This is the same admittance as the circuit shown in Fig. 4.29. Thus, the input admittance is same as a capacitor in parallel with a series combination of a negative capacitor and a negative resistor. If a third network consisting of a capacitor of size C_1 and a resistor of size R_1, in series, was connected to the gate of the source-follower as shown in Fig. 4.30, then the negative elements would be cancelled. The resulting input admittance would then simply be C_2 as given in (4.216). In this case (4.197) becomes

$$\frac{v_{g1}}{i_{in}} = \frac{1}{G_{in} + s\left(C'_{in} + \dfrac{C_{gs1}C_s}{C_{gs1} + C_s}\right)} \tag{4.217}$$

and (4.199) becomes

$$A(s) = \frac{v_{out}}{i_{in}} = R_{in}\left(\frac{g_{m1}}{g_{m1} + G_{s1}}\right)\frac{\left(1 + s\dfrac{C_{gs1}}{g_{m1}}\right)}{\left(1 + \dfrac{s}{p_1}\right)\left(1 + \dfrac{s}{p_2}\right)} \tag{4.218}$$

where

$$p_1 = \frac{G_{in}}{C'_{in} + \dfrac{C_{gs1}C_L}{C_{gs1} + C_L}} \cong \frac{G_{in}}{C'_{in} + C_{gs1}} \tag{4.219}$$

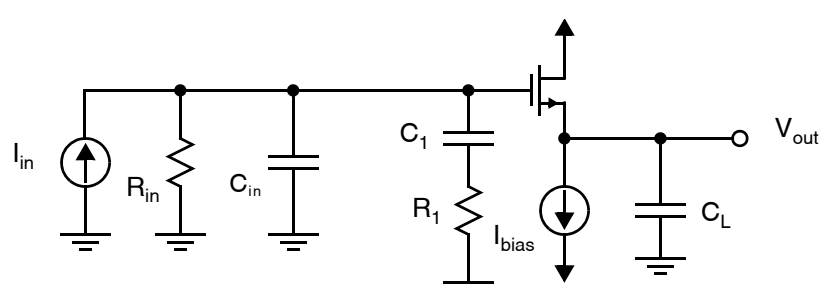


Fig. 4.30 Adding a compensation network (C_1 and R_1) to compensate for the negative components of the admittance looking into the gate of the source-follower.

$$p_2 = \frac{g_{m1} + G_{s1}}{C_{gs} + C_L} \cong \frac{g_{m1} + G_{s1}}{C_L} \tag{4.220}$$

The approximation is accurate when $C_s \gg C_{gs1}$. Irrespective of the approximation, the poles are now guaranteed real and no overshoot will occur.

EXAMPLE 4.17

Using the same parameters as in Example 4.16, find the compensation network and the resulting first and second poles of the source follower of Fig. 4.26.

Solution

Using (4.216), we have

$$C_1 \cong \frac{g_{m1} C_{gs1} C_s}{(g_{m1} + G_{s1})(C_{gs1} + C_s)} = 2.37\,\text{fF} \tag{4.221}$$

and

$$R_1 \cong \frac{(C_{gs1} + C_s)^2}{C_{gs1} C_s g_{m1}} \cong 93\text{k}\Omega \tag{4.222}$$

The capacitor is a small but reasonable value to be realized on chip. The resistor could be realized by a MOS transistor biased in the triode (i.e., linear) region. Assuming the compensation network is used, the poles of the transfer function would then become

$$p_1 \cong \frac{G_{in}}{C_{gs1} + C_{gd1}} = 2\pi \times 2.39\text{GHz} \tag{4.223}$$

and

$$p_2 = \frac{g_{m1} + G_{s1}}{C_{gs} + C_L} = 2\pi \times 0.72\text{GHz} \tag{4.224}$$

Finally, it should be mentioned here that if the source-follower buffer is intended to be used in an opamp (and thus feedback will be placed around the buffer), and if the resonant frequency of the source-follower is substantially greater than the unity-gain frequency of the amplifier, then the overshoot can be tolerated and no compensation network is necessary.

4.5 DIFFERENTIAL PAIR

4.5.1 High-Frequency T Model

Similar to low-frequency analysis, there exists a *T model* for high-frequency modelling that sometimes results in simpler analyses and greater insight, especially when the gain is not large. Consider the small-signal T model for a MOSFET shown in Fig. 4.31. The T model significantly simplifies the analysis of differential-pair-based amplifiers especially when the transistor output resistors can be ignored.

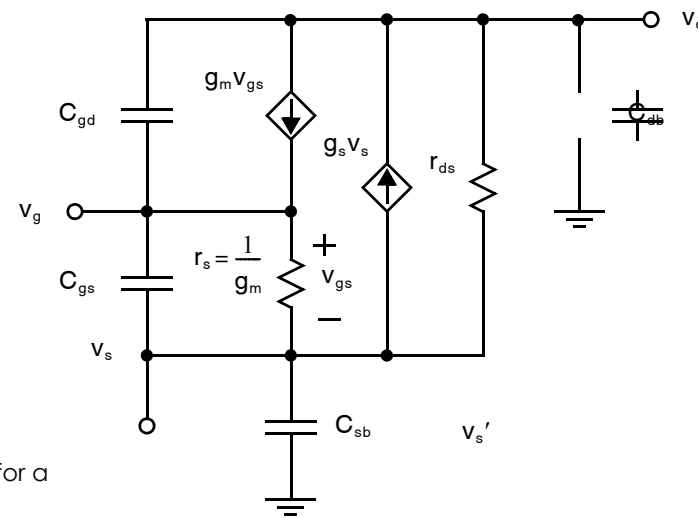


Fig. 4.31 A high-frequency T model for a MOSFET in the active region with the body terminal at small-signal ground.

4.5.2 Symmetric Differential Amplifier

A symmetric differential amplifier with resistive loads is shown in Fig. 4.32. A small-signal model of this amplifier using the T-model is shown in Fig. 4.33. Here we have ignored both the transistor drain-source resistances in order to simplify the circuit.

Key Point: *The analysis of a symmetric differential pair may be simplified to that of the half-circuit, which is simply a common-source amplifier.*

Assuming that the transistors are matched, then the circuit is perfectly symmetric. If one also assumes that $v_{in}^- = -v_{in}^+$, then based on the symmetry, the node voltage v_s will never change similar to if it was connected to a small-signal ground. Indeed, if the node v_s was connected to a small-signal ground, then circuit will operate identically. This allows one to ignore C_{sb1} and C_{sb2} (not included in Fig. 4.33) and simplify the analysis to that of the *half-circuit* which is identical to the common-source amplifier studied previously. Given this equivalence, and

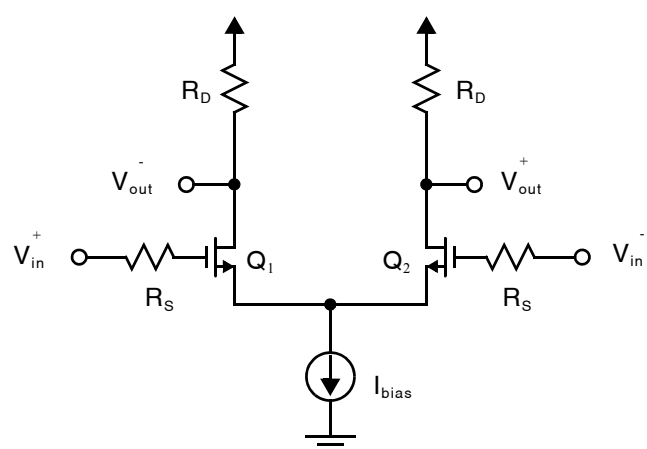


Fig. 4.32 A symmetric MOS differential amplifier.

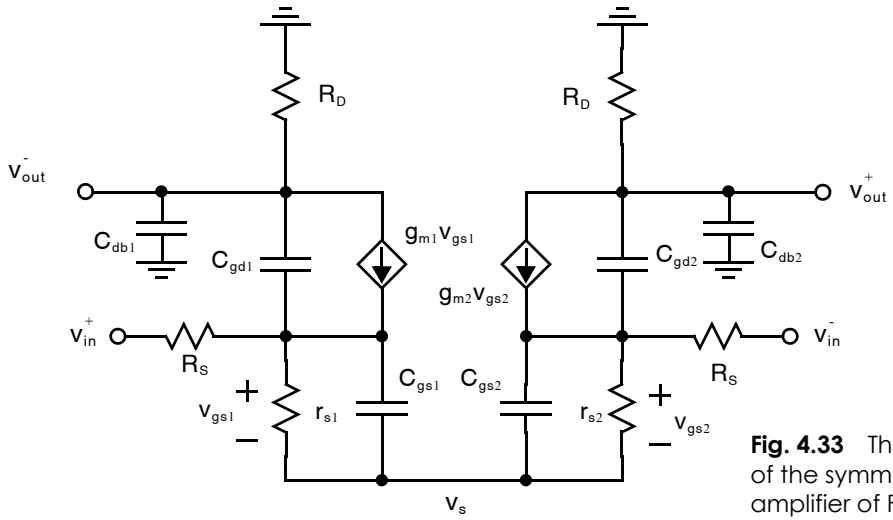


Fig. 4.33 The small-signal model of the symmetric differential amplifier of Fig. 4.32.

assuming the time constant associated with C_{db} is negligible, one can immediately estimate the amplifier bandwidth.

$$\omega_{-3dB} \cong \frac{1}{R_s[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)]} \tag{4.225}$$

4.5.3 Single-Ended Differential Amplifier

Consider next the frequency response of the single-ended output differential amplifier shown in Fig. 4.34. Before we analyze this amplifier for its frequency response, note that this configuration can be considered a two-stage amplifier comprising an source follower followed by a common-gate stage. The small-signal model for this amplifier is shown in Fig. 4.34(*b*). Note both the similarities and the differences of the circuits shown in Fig. 4.33 and Fig. 4.34(*b*). The differences, though they appear minor result in a significantly larger bandwidth (and approximately half the gain).

Before starting the analysis, note that between the gate of Q_1 and ground are two similar networks, namely r_{s1} in parallel with C_{gs1}, and r_{s2} in parallel with C_{gs2}. If we assume the bias current I_{bias} is split equally between Q_1 and Q_2, we may simplify the circuit into the equivalent small-signal model shown in Fig. 4.35. This simplification assumes the transistor parameters are matched which allows one to drop the subscripts. Next notice that the left-most dependant current source is proportional to the voltage across it; this makes it equivalent to a negative resistor (given the direction of the current source) of size $-2/g_m$. We can simplify the circuit even more by noting that the negative resistance $-2/g_m$ perfectly cancels out the parallel resistance $2r_s = 2/g_m$. The final simplified circuit is now shown in Fig. 4.36. The transfer function is now easily found almost by inspection to be

$$H(s) = \frac{-\frac{g_m}{2}R_D}{\left(1 + sR_s\left(\frac{C_{gs}}{2} + C_{gd1}\right)\right)(1 + sC_{gd2}R_D)} \tag{4.226}$$

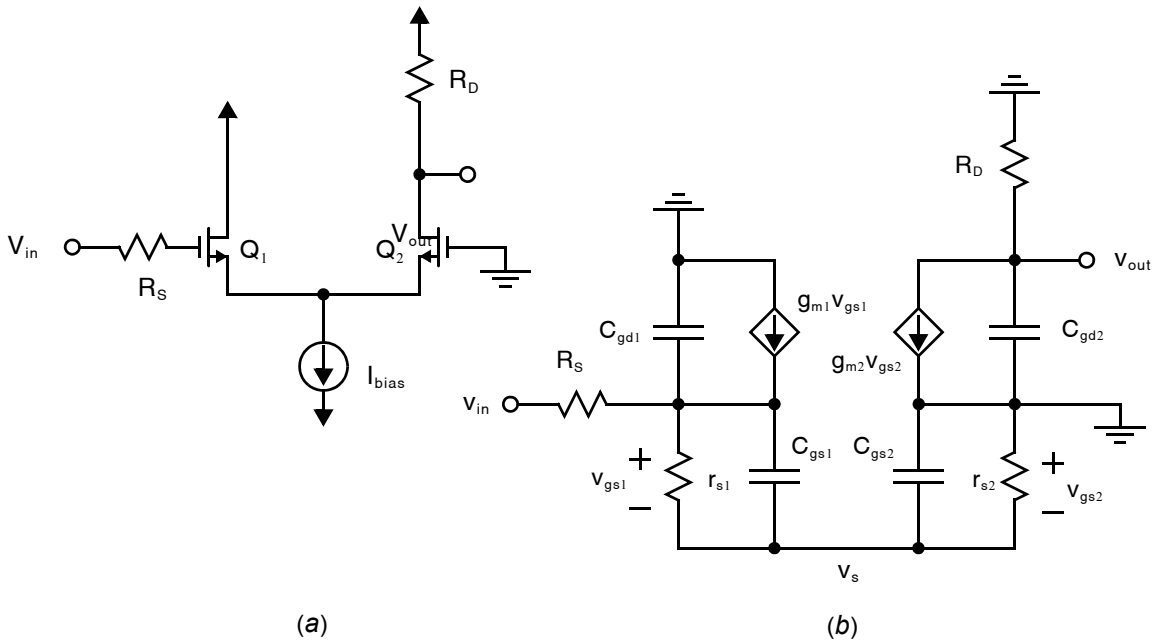


Fig. 4.34 A differential pair with single-ended output: (*a*) complete circuit; (*b*) small-signal equivalent.

The first pole frequency due to the time constant at the gate of Q_1 is given by

$$\omega_{p1} = \frac{1}{R_s\left(\frac{C_{gs}}{2} + C_{gd1}\right)} \tag{4.227}$$

Note that there is no Miller-multiplication term (i.e., $1 + A_0$) multiplying the capacitance C_{gd1}. For this reason, the compound stage is must faster than a common-source stage, although it has similar gain.

4.5.4 Differential Pair with Active Load

We next consider the high-frequency analysis of a MOS differential pair with an active current-mirror load. Described first in Section 3.8 and shown in Fig. 3.19, this stage is often used as the input to a classic two-stage

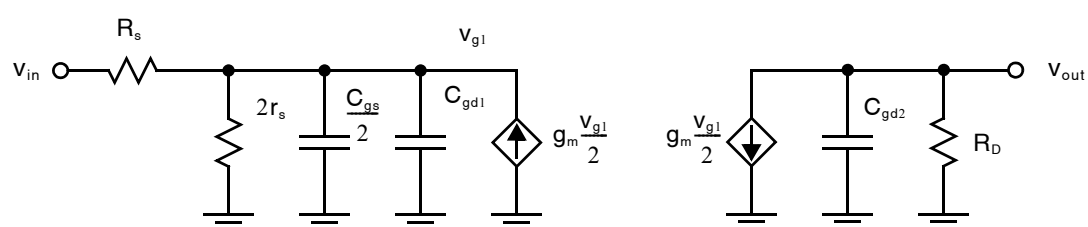


Fig. 4.35 A simplified and equivalent small-signal model of the amplifier of Fig. 4.34.

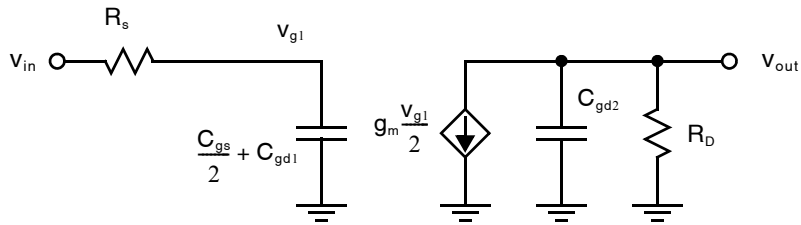


Fig. 4.36 A further simplified small-signal model of the amplifier of Fig. 4.34.

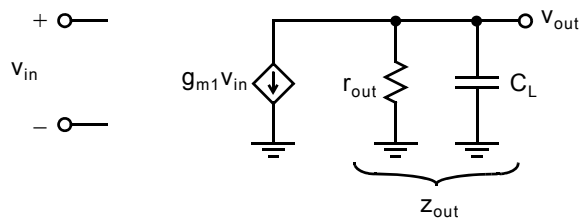


Fig. 4.37 A small-signal model for the differential-input amplifier.

operational amplifier circuit. In that case, it drives a significant capacitive load, C_L. Assuming that load is dominant, we can easily modify the low-frequency small-signal analysis of this circuit performed in Section 3.8 by substituting r_{out} with z_{out},

$$A_v = \frac{v_{out}}{v_{in}} = g_{m1} z_{out} \tag{4.228}$$

where $z_{out} = r_{out} \parallel 1/(sC_L)$. Thus, for this differential stage, the very simple model shown in Fig. 4.37 is commonly used. This model implicitly assumes that only the capacitance at the output node is significant and the parasitic capacitances at the node at the sources of $Q_{1,2}$ and the node at the gates of $Q_{3,4}$ may be ignored. This assumption is usually justified, because the small-signal resistance at the output node, r_{out}, is much larger than the small-signal resistances at the other nodes which are generally $\approx 1/g_m$. Also, the capacitance at the output node, C_L, is usually larger than the parasitic capacitances at the other nodes. Hence, the time-constant associated with the other capacitances will be much smaller than the time-constant associated with C_L and the –3dB bandwidth is well approximated by

$$\omega_{-3dB} \cong \frac{1}{r_{out} C_L} = \frac{1}{(r_{ds2} \parallel r_{ds4}) C_L} \tag{4.229}$$

However, when high-frequency effects are important (which may be the case when compensating an opamp to guarantee stability), then this assumption may not be justified.

4.6 KEY POINTS

- The transfer functions in analog circuits throughout this text: a) are rational with $m \leq n$ b) have real-valued coefficients, a_i and b_i c) have poles and zeros that are either real or appear in complex-conjugate pairs. Furthermore, if the system is stable, d) all denominator coefficients $b_i > 0$ e) the real part of all poles will be negative. [p. 145]

- When a linear circuit is excited with a sinusoid, the output will be a sinusoid at the same frequency. Its magnitude will equal to the input magnitude multiplied by the magnitude response $|H(\omega_{in})|$. The phase difference between the output and input sinusoids will equal the phase response $\angle H(\omega_{in})$. Magnitude responses are often expressed in units of decibels, $20\log_{10}|H(\omega)|$ dB. [p. 146]
- For a first-order lowpass transfer function with dc gain $A_0>>1$, the unity gain frequency is $\omega_{ta}\approx A_0\omega_{-3dB}$ and $\angle A(\omega_{ta})\approx 90°$. [p. 154]
- Transfer functions with only real-valued poles and numerator order zero have monotonically decreasing magnitude and phase response plots with each pole contributing an additional –20 dB/decade slope in the magnitude response, and an additional –90° phase shift. [p. 157]
- When a circuit has one or more capacitors with relatively very large value, one may consider those capacitors short-circuited to see how the circuit behaves at higher frequencies. Similarly, when one or more capacitors have relatively very small value, the circuit may be considered with those capacitors open-circuited to understand its operation at lower frequencies. [p. 161]
- The frequency-response of the common-source amplifier has 2 poles and 1 zero. The complexity of the analysis illustrates the importance of making approximations and simplifications in the high-frequency analysis of analog circuits. [p. 169]
- When a capacitor connects the input and output of a high-gain inverting amplifier, it appears much larger at the input than it really is. This Miller effect is often responsible for the dominant pole of such amplifiers and, hence, is useful for obtaining a quick estimate of bandwidth. [p. 172]
- The method of zero-value time-constants estimates the dominant pole and, hence, bandwidth of complex circuits by analyzing several simpler circuits at dc. [p. 173]
- All small-signal capacitances in the common-gate amplifier are (small-signal) grounded. Hence there is no Miller effect and it can provide high-frequency performance superior to that of the common-source amplifier, although with much lower input resistance. [p. 181]
- The cascode gain stage uses a common-gate transistor Q_2 to reduce the V_{DS} variations on a common-source transistor Q_1. The result is high output resistance providing potentially high gain, a reduced Miller effect, and reduced short-channel effects. However, the available output voltage swing is reduced. [p. 182]
- The unity-gain frequency of any feedback amplifier that uses a cascode gain stage with a dominant output pole is limited by the unity-gain frequency of the cascode transistor, Q_2. [p. 187]
- Source follower circuits can exhibit large amounts of overshoot and ringing under certain conditions. When designing source-followers, the recommended procedure is to check to see if the poles are complex using either (4.203) or a SPICE transient analysis to look for overshoot. When the poles are complex, then increase the load and/or input capacitor or, alternatively, add the compensating network shown in Fig. 4.30. [p. 192]
- The analysis of a symmetric differential pair may be simplified to that of the half-circuit, which is simply a common-source amplifier. [p. 194]

4.7 REFERENCES

P. R. Gray, P. J. Hurst, S. H. Lewis, R. G. Meyer. *Analysis an d Design o f An alog Integrated Cir cuits*, 5th ed, Wiley & Sons, Hoboken, NJ, 2009.

G. James. *Advanced Modern Engineering Mathematics*, 3rd ed., Pearson Education Ltd., Harlow, England, 2004.

A. Sedra, K. C. Smith. *Microelectronic Circuits*, 6th ed., Oxford University Press, Oxford, England, 2009.

4.8 PROBLEMS

4.8.1 Section 4.1: Frequency Response of Linear Systems

4.1 Express the following voltage gains in decibels:

a. $A_v = 42$ V/V

b. $A_v = 0.3$ V/V

c. $A_v = -1500$ V/V

4.2 A linear system has a transfer function

$$H(s) = \frac{s}{s^2 + 5s + 50}$$

a. Sketch a Bode plot of $H(s)$.

b. What is the maximum gain $|H(\omega)|$ expressed in dB? At what frequency does it occur?

4.3 Consider the CR network in Fig. P4.3 with $R = 2\ k\Omega$ and $C = 15$ pF.

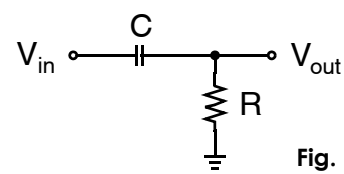


Fig. P4.3

a. Write the transfer function $V_{out}(s)/V_{in}(s)$.

b. Sketch the step response of the circuit.

4.4 Two linear systems, $H_1(s)$ and $H_2(s)$ are connected in series so that the output of $H_1(s)$ is connected to the input of $H_2(s)$. Their respective Bode plots are sketched in Fig. P4.4.

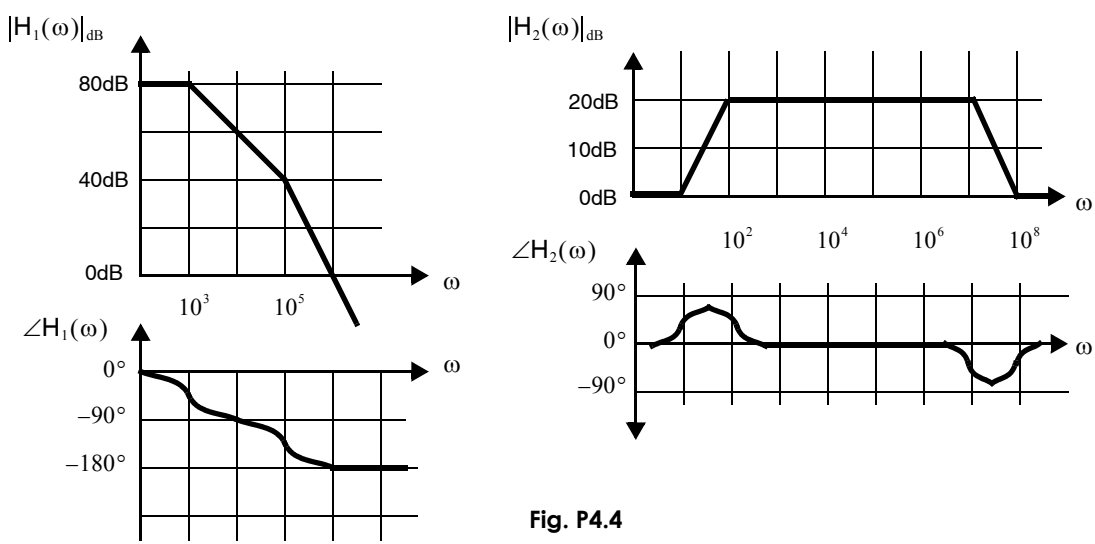


Fig. P4.4

a. Sketch the bode plot of the combined system, $H_1(s) \cdot H_2(s)$.

b. Estimate all of the pole and zero locations.

4.5 For the second-order system $H_1(s)$ whose Bode plot is sketched in Fig. P4.4, how long does it take for the step response to settle within 1% of its final value?

4.6 In Fig. P4.6, $C_1 = 30$ pF and $C_2 = 100$ fF. Draw an approximately-equivalent circuit for low-medium frequencies. At what frequency does this approximate circuit cease to be valid? Compare the transfer functions, Bode plots, and step responses of the approximate and exact circuits.

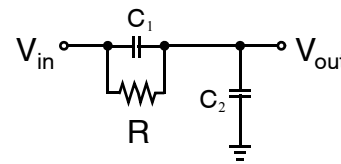


Fig. P4.6

4.7 Write a second-order lowpass transfer function with a dc gain of 12 dB, a Q-factor of 0.2, and $\omega_0 = 2\pi \cdot 3$ MHz.

a. What are the pole locations of this transfer function?

b. Sketch a Bode plot of the transfer function.

c. How long will it take for the step response to settle within 5% of its final value?

4.8 Write a second-order lowpass transfer function with a dc gain of 23 dB, a Q-factor of 2, and $\omega_0 = 2\pi \cdot 20$ MHz.

a. What are the pole locations of this transfer function?

b. How much overshoot does the step response exhibit? How long will it take to settle within 1% of its final value?

4.8.2 Section 4.2: Frequency Response of Elementary Transistor Circuits

4.9 Estimate the bandwidth of the common-source amplifier in Example 4.11.

4.10 Find the zero-value time constants of the common-source amplifier in Example 4.12. Confirm that the output time-constant dominates the amplifier's bandwidth.

4.11 Consider the source-degenerated common-source amplifier shown in Fig. P4.11 neglecting the transistor parasitic capacitances.

a. Draw an equivalent small-signal circuit at dc. What is the dc small-signal gain, v_o/v_i?

b. Draw an equivalent small-signal circuit at frequencies high enough that C_s may be considered a short circuit. What is the small-signal gain v_o/v_i at these frequencies?

c. Make a sketch of the magnitude response $|v_o/v_i|$ versus frequency covering the range of frequencies analyzed in parts (a) and (b) above.

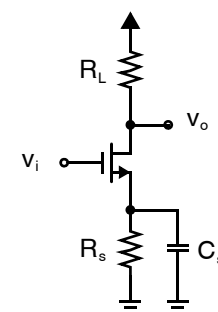


Fig. P4.11

4.12 Consider the common source amplifier in Fig. P4.12 driving a capacitive load of $C_L = 0.5$ pF. Assume $L = L_{min}$ and the gate dc bias voltage V_G is set to keep Q_1 in saturation. What current I_D and device width W are required to maximize the small-signal dc gain $|v_o/v_i|$ while maintaining a 3-dB bandwidth of at least 10 MHz?

a. Use the NMOS device parameters for the 0.18- μm CMOS process in Table 1.5.

b. Use the NMOS device parameters for the 45-nm CMOS process in Table 1.5.

Compare your results with SPICE.

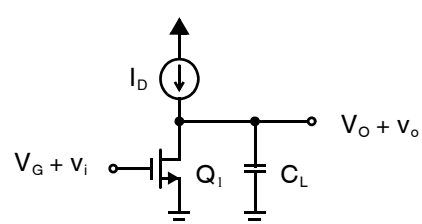


Fig. P4.12

4.13 Again consider the common source amplifier in Fig. P4.12, this time driving a load capacitance of $C_L = 0.4$ pF. What is the minimum current I_D required to achieve a 3-dB bandwidth of 3 MHz using the 0.35-μm CMOS process in Table 1.5? Compare your result with SPICE.

4.14 The common source amplifier in Fig. P4.14 drives a load capacitance of $C_L = 25$ pF. What is the maximum unity-gain frequency achievable with $I_D = 0.8$ mA?

a. Use the device parameters for the 0.18-μm CMOS process in Table 1.5.

b. Use the device parameters for the 45-nm CMOS process in Table 1.5.

Compare your results with SPICE.

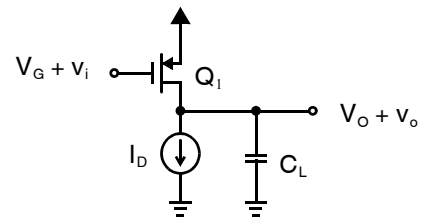


Fig. P4.14

4.15 A current source with a purely capacitive source impedance $Z_s = 1/(sC_s)$ drives a NMOS common-gate amplifier. Select the common-gate transistor size W/L so that the input time constant is 1 ns with minimal bias drain current I_D when $C_s = 5$ pF. Use the device parameters for the 45-nm CMOS process in Table 1.5.

4.16 A common-gate amplifier is to be designed to have an input resistance $r_{in} = 50\ \Omega$ using the 0.35-μm CMOS process in Table 1.5. and $I_D = 2$ mA. Select W/L of the common-gate transistor to minimize the resulting input time constant. Estimate the input time constant.

4.17 The circuit in Fig. P4.17 is referred to as an "active inductor."

a. Find an expression for the small-signal input impedance as a function of frequency in terms of the circuit parameters R, C, g_m, etc. assuming $C \gg C_{gd}, C_{sb}$.

b. Over what range of frequencies is Z_{in} well modeled by an inductor?

c. What is the value of the effective inductance in terms of the circuit parameters?

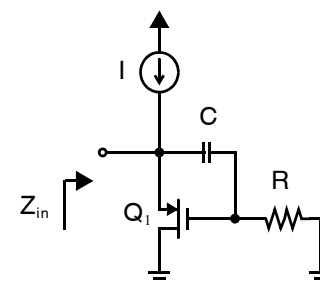


Fig. P4.17

4.18 Find and expression for the small-signal frequency response $(i_{out}/i_{in})(\omega)$ for the simple NMOS current mirror shown in Fig. P4.18. Find a simplified approximate expression for the 3-dB bandwidth of the current mirror assuming:

a. the load capacitor C_L is very large;

b. the load capacitor C_L is zero.

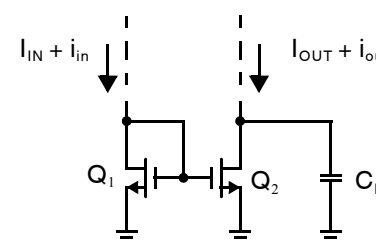


Fig. P4.18

4.8.3 Section 4.3: Cascode Gain Stage

4.19 For the telescopic cascode amplifier in Fig. 4.22(*a*), $I_{bias} = 300\ \mu A$ and both devices are sized with $(W/L) = 10\ \mu m/0.18\ \mu m$. Select the dc bias voltage at V_{in} and V_{bias} to maximize the output swing available at V_{out} while keeping Q_1 and Q_2 in active mode. Use the transistor parameters for the 0.18-μm CMOS process in Table 1.5.

4.20 Consider the folded cascode amplifier in Fig. 4.22(*b*) with $I_{bias1} = 400\ \mu A$, $I_{bias2} = 150\ \mu A$ and both devices sized $(W/L) = 10\ \mu m/0.18\ \mu m$. The dc bias voltage $V_{bias} = 0.8$ V. What is the maximum voltage that can appear at V_{out} so that Q_2 remains in active mode? Use the transistor parameters for the 0.18-μm CMOS process in Table 1.5.

4.21 The folded cascode amplifier in Fig. 4.22(*b*) has $g_{m1} = g_{m2} = 0.5$ mA/V and $V_{eff,1} = V_{eff,2} = 0.2$ V. Determine I_{bias1} and I_{bias2}. Determine the transistor widths assuming $L = L_{min}$ in each of the following cases:

a. Using the 0.35-μm CMOS device parameters in Table 1.5.

b. Using the 0.18-μm CMOS device parameters in Table 1.5.

4.22 The telescopic cascode amplifier in Fig. 4.22(*a*) has $I_{bias} = 400\ \mu A$ and both devices are sized with $(W/L) = 20\ \mu m/0.18\ \mu m$. The bias voltages are selected to keep Q_1 and Q_2 in active mode. The load capacitance is $C_L = 6$ pF. Estimate the dc gain and circuit bandwidth using the transistor parameters for the 0.18-μm CMOS process in Table 1.5.

4.23 You are to design an amplifier to provide a gain of 8 using the 0.35-μm CMOS process in Table 1.5. The input voltage source resistance is $50\ k\Omega$ and the load is $4\ k\Omega$. Assume the supply voltage is 3 V and the power consumption is to be 12 mW.

a. If a NMOS common-source amplifier is used, what W/L will maximize the bandwidth? Estimate the bandwidth.

b. If a NMOS telescopic cascode amplifier is used, and assuming $(W/L)_1 = (W/L)_2$, size the transistors to maximize bandwidth. Estimate the resulting bandwidth and compare it to the result in part (a).

c. Repeat part (b), but this time design a PMOS cascode amplifier. How does your bandwidth estimate compare to that of part (b)?

4.8.4 Section 4.4: Source-Follower Amplifier

4.24 Perform a zero-value time constant analysis on the common-source amplifier in Fig. 4.28. What are the required conditions for the time constant associated with C_s to be dominant? Provide an approximate expression for the bandwidth in this case.

4.25 The source-follower shown in Fig. 4.26 has $I_{bias} = 1.5$ mA, $R_{in} = 5\ k\Omega$, $C_{in} = 30$ fF, a load capacitance of $C_L = 8$ pF and $(W/L) = (100\ \mu m/0.35\ \mu m)$. Estimate the bandwidth using the 0.35-μm CMOS device parameters in Table 1.5. Compare your result with SPICE. How does this compare with the bandwidth that would be obtained if C_L were driven directly with the source whose internal resistance is R_{in}?

4.26 A source-follower is to be designed to drive a load capacitance of $C_L = 4$ pF with 500 MHz bandwidth using the 0.18-μm CMOS process described in Table 1.5. What is the minimum value of I_{bias} if

a. $V_{eff} = 150$ mV?

b. $V_{eff} = 250$ mV?

c. Select the corresponding device size (W/L) in each case.

4.8.5 Section 4.5: Differential Pair

4.27 Consider the amplifier in Fig. 4.32 with $I_{bias} = 300\ \mu A$, $(W/L)_{1,2} = 12\ \mu m/0.2\ \mu m$, and $R_S = 10\ k\Omega$. Select R_D to provide a gain of 3 and estimate the resulting amplifier bandwidth using the 0.18-μm CMOS device parameters in Table 1.5.

4.28 Repeat Problem 4.27, but this time add load capacitances of $C_L = 6$ pF between V_{out}^+ and ground and between V_{out}^- and ground.

4.29 Design the CMOS differential pair circuit in Fig. P4.29 to have a differential gain of 3 V/V and a 3-dB bandwidth of 700 MHz when $C_L = 1.5$ pF. The transistors should be biased with $V_{eff} = 0.2$ V and should have gate lengths of 0.2 μm. Using the 0.18-μm CMOS parameters in Table 1.5, determine the widths of the transistors and the values of R_D and I_{bias}.

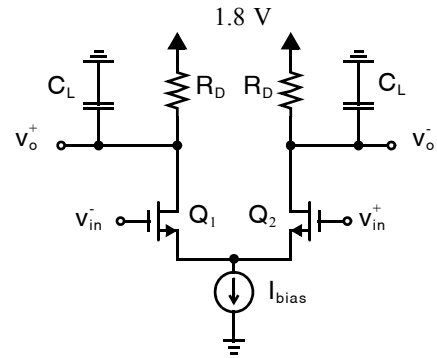


Fig. P4.29

4.30 The circuit in Fig. P4.30 is perfectly symmetric and the small signal input voltages are balanced so that $V_{in}^{\pm} = V_{IN} \pm v_{in}/2$. The bias voltages V_b and V_{IN} are chosen to ensure all transistors remain in active mode. What is the equivalent half-circuit in this case? Estimate the amplifier's 3-dB bandwidth.

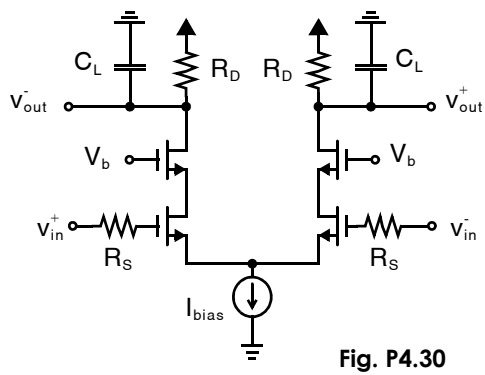


Fig. P4.30

4.31 Find an approximate expression for the bandwidth of the differential amplifier in Fig. P4.31. All transistors are sized identically. Assume the bias voltage V_b is chosen so that all transistors remain in active mode. How does the bandwidth compare with that of the conventional resistively-loaded differential pair in Fig. 4.32?

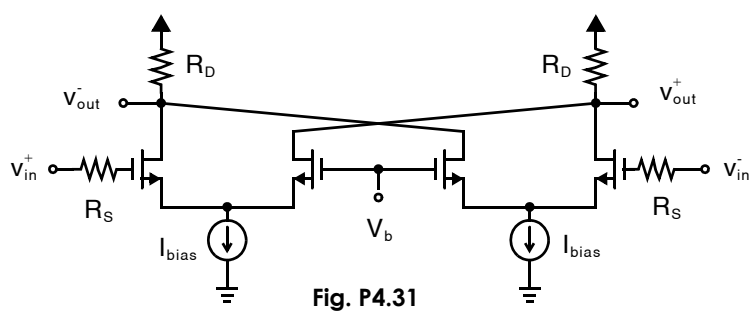


Fig. P4.31

4.32 Design the amplifier of Fig. P4.29 using the 0.18-μm CMOS process in Table 1.5 with $I_{bias} = 400$ μA to provide a gain of 20 dB. Take all gate lengths $L = 0.3$ μm. Estimate the resulting bandwidth when driving a 2-pF capacitive load.

CHAPTER

5 Feedback Amplifiers

Feedback is a topic of paramount importance in analog design. It is a key concept that enables us to construct almost ideal analog circuits from the poorly controlled and very nonideal transistors readily available to us. By doing so, it facilitates the division of complex systems into smaller subblocks that may be designed independently.

5.1 IDEAL MODEL OF NEGATIVE FEEDBACK

We begin by considering the simplified model of a very general negative feedback system in Fig. 5.1. Its analysis finds overlap in introductory control theory and is applicable to a wide variety of systems, including analog integrated circuits. Later in this chapter, we refine our analysis making it progressively better suited to the circuits of most interest to us.

5.1.1 Basic Definitions

The linear time-invariant negative feedback system in Fig. 5.1 includes an amplifier, A, and some feedback circuitry whose gain is β. Also labelled in Fig. 5.1 are the input signal, u, the output signal, y, the feedback signal, v, and the difference signal,

$$x = u - v \tag{5.1}$$

Physically, each of these signals may be small-signal node voltages or branch currents.[1]

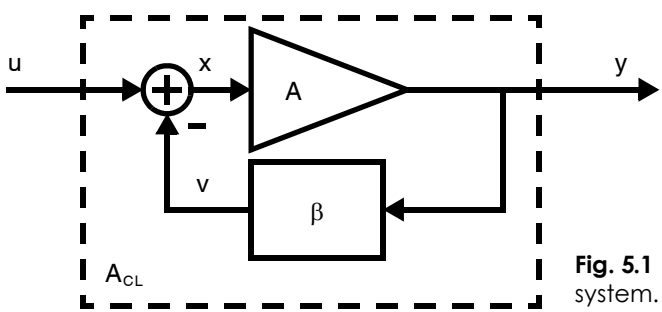


Fig. 5.1 Ideal model of a negativefeedback system.

1. In some cases, it may be difficult to identify precisely which circuit variables correspond to the signals in Fig. 5.1. The model is applied to a few practical examples in Section 5.4.

Clearly,

$$v = A\beta x \equiv Lx \tag{5.2}$$

where we have defined the open-loop gain, or simply "loop gain,"

$$L \equiv A\beta \tag{5.3}$$

Since $x = u - v$ in Fig. 5.1, they must all have the same units. Hence, the loop gain L is always dimensionless.

If we consider the entire feedback system as a unified single-input single-output block, its input-output relationship is derived as follows:

$$x = u - Lx \Rightarrow x = \frac{u}{1+L} \tag{5.4}$$

$$y = Ax = \frac{A}{1+L}u = \frac{A}{1+A\beta}u \tag{5.5}$$

The "closed-loop" gain follows from (5.5),

$$A_{CL} \equiv \frac{y}{u} = \frac{A}{1+A\beta} = \frac{A}{1+L} \tag{5.6}$$

and is not to be confused with the loop gain in (5.3).

5.1.2 Gain Sensitivity

The gain of amplifier A depends upon the small-signal parameters of its constituent transistors (i.e., their transconductances, g_m, and output resistances, r_{ds}). These in turn are subject to large variations over temperature and from one sample of the circuit to another. Fortunately, the value of the closed-loop gain, A_{CL}, is largely immune to these variations so long as the loop-gain, L, is much larger than unity.

To quantify this, consider the effect of a change in the amplifier gain from A to $A + \Delta A$ while keeping the value of β constant. The resulting change in closed-loop gain is

$$\begin{aligned} \Delta A_{CL} &= \frac{A + \Delta A}{1 + (A + \Delta A)\beta} - \frac{A}{1 + A\beta} \\ &= \frac{(A + \Delta A)(1 + A\beta) - A[1 + (A + \Delta A)\beta]}{[1 + (A + \Delta A)\beta](1 + A\beta)} \\ &= \frac{\Delta A}{(1 + L)^2 + \Delta A\beta} \end{aligned} \tag{5.7}$$

Hence, if $L \gg 1$, large changes in the amplifier gain, ΔA, result in very little change in A_{CL}. In other words, negative feedback enables high accuracy in the gain of the overall amplifier, even when there are large variations in the transistor parameters of the transistors providing the gain.

EXAMPLE 5.1

Consider the negative feedback system pictured in Fig. 5.1 where the amplifier inputs and outputs are voltages, so that both A and β are dimensionless. If the amplifier gain is 80 dB and $\beta = 0.2$, what is the resulting closed-loop gain, A_{CL}? If A decreases by an order magnitude, what is the resulting change in A_{CL}?

Solution

A gain of 80 dB corresponds to a value of $A = 10^4$. Plugging this value into (5.6) results in a closed-loop gain of

$$A_{CL} = \frac{10^4}{1 + 0.2 \cdot 10^4} = 4.9975$$

If the gain decreases by an order of magnitude to 60 dB ($A = 10^3$) the closed-loop gain becomes

$$A_{CL} = \frac{10^3}{1 + 0.2 \cdot 10^3} = 4.9751$$

representing a change of less than 0.5%.

Key Point: *When* $L>>1$*, the closed-loop system has an overall gain inversely proportional to the gain of the feedback circuit and highly insensitive to the gain of the amplifier,* A.

The insensitivity of A_{CL} to changes in A is also evident if one rewrites (5.6) substituting $A = L/\beta$.

$$A_{CL} = \frac{1}{\beta} \cdot \frac{L}{(1 + L)} \tag{5.8}$$

The closed loop gain is equal to a constant, the "*desired gain*" $1/\beta$, multiplied by an error term that is close to unity as long as $L \gg 1$. In the limit of very large loop gain, the closed loop gain approaches the desired gain.

$$A_{CL} \cong \frac{A}{L} = \frac{1}{\beta} \tag{5.9}$$

Hence, to obtain high accuracy in the closed-loop gain it is necessary to have high accuracy in the gain of the feedback circuit.

EXAMPLE 5.2

Repeat Example 5.1, but this time consider a change in β from 0.2 to 0.21 while A remains constant at 80 dB.

Solution

With $\beta = 0.21$, the closed-loop gain becomes

$$A_{CL} = \frac{10^4}{1 + 0.21 \cdot 10^4} = 4.7596$$

This is approximately 5% lower than the result obtained in Example 5.1 for $\beta = 0.2$, a percentage change commensurate with the change in β.

5.1.3 Bandwidth

The amplifier gain A is frequency-dependent, generally decreasing in magnitude at higher frequencies, as depicted in Fig. 5.2. However, as long as the loop gain remains much greater than unity (i.e. $A(\omega) \gg 1/\beta$) the closed-loop gain will remain approximately $A_{CL} \cong 1/\beta$. At very high frequency, eventually $A(\omega) \ll (1/\beta) \Rightarrow L \ll 1$ and (5.6) simplifies to $A_{CL} \cong A(\omega)$. A very rough estimate of the bandwidth of the closed-loop amplifier is therefore provided by the transition between these zones; that is, the frequency at which $|A(\omega)| = (1/\beta)$, or equivalently where $|L(\omega)| = 1$. It is clear from Fig. 5.2 that this closed-loop bandwidth will generally be much higher than the bandwidth of $A(\omega)$.

> ***Key Point:*** *A rough estimate of the bandwidth of a feedback amplifier is given by the frequency where* $|L(\omega)| = 1$*, which is much higher than the bandwidth of* $A(\omega)$.

The sketch in Fig. 6.2, is predicated on the assumption that β remains constant over the entire frequency range under consideration. Hence, to realize a closed-loop amplifier with high bandwidth, it is necessary for the feedback network to have high bandwidth, but not the amplifier, A.

5.1.4 Linearity

The model in Fig. 5.1 is completely linear and, hence, is only valid so long as the transistors in the amplifier, A, are exposed to small voltage and current signals. Fortunately, by applying negative feedback with large loop gain, the signal that appears at the amplifier input, x, is a strongly attenuated version of the input to the overall system, u. With $L \gg 1$, (5.4) becomes

$$x \cong \frac{u}{L} \tag{5.10}$$

This means that even when large signals are present at u, the amplifier A is exposed to only very small voltage and current signals at its input, and can therefore operate linearly. The feedback circuit, however, must operate linearly even when exposed to large signals. Its input is the amplifier output y which is A times greater than x.

> ***Key Point:*** *With feedback, an amplifier is exposed to signals that are approximately* L *times smaller than would be the case without feedback, making it much more linear.*

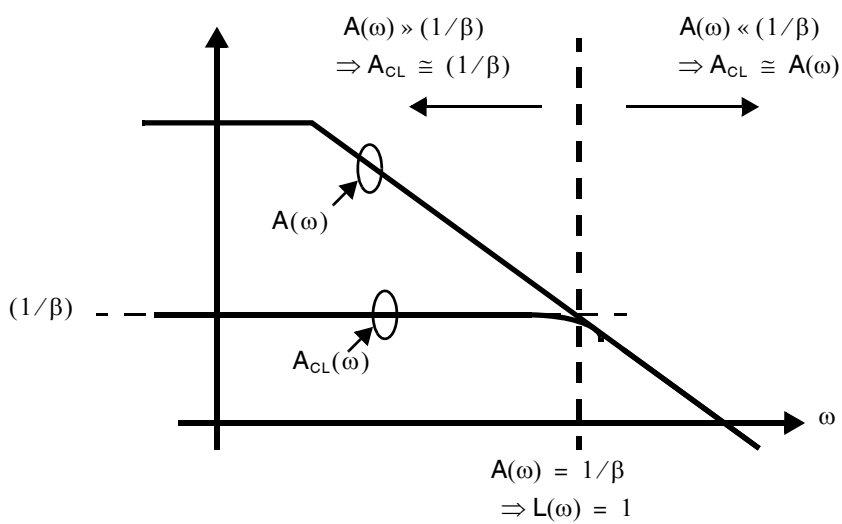


Fig. 5.2 Bandwidth of a closed-loop amplifer.

5.1.5 Summary

The preceding subsections show that amplifiers with an accurate gain of $A_{CL} = 1/\beta$, high bandwidth, and high linearity may be realized by using negative feedback as long as:

1) the feedback circuit has an accurately controlled gain, β, high bandwidth, and high linearity
2) the loop gain $L = A\beta$ is large
3) the feedback loop is stable

The conditions on the feedback circuit (#1) are most easily satisfied by constructing it out of passive components and/or transistors acting as simple switches, thus avoiding the variability, parasitic poles, and inherent nonlinearity of circuits with transistors in the active mode. Often, the gain is dependent only on the *ratios* of passive component values, which in integrated circuits are more accurate than their absolute values. Condition #2 is satisfied by designing an amplifier with large gain, $A \gg 1/\beta$. However, this is virtually the only specification on the amplifier since feedback greatly relaxes the requirements on its accuracy, bandwidth, and linearity.

Condition #3, concerning stability of the loop, is the topic of the next subsection.

5.2 DYNAMIC RESPONSE OF FEEDBACK AMPLIFIERS

The reader may have already made use of feedback amplifiers without explicitly considering their stability. However, such methods can fail to predict critical circuit behavior.

EXAMPLE 5.3

The noninverting amplifier in Fig. 5.3(*a*) is often analyzed assuming the opamp's differential input is zero. Hence, the negative amplifier terminal voltage is $v_- = v_{in}$. Further assuming the amplifier has infinite input impedance, the output voltage may be related to the input by a nodal equation at v_-.

$$\begin{aligned} v_{in}\left(\frac{1}{R_1}+\frac{1}{R_2}\right) - \frac{v_{out}}{R_2} &= 0 \\ \Rightarrow \frac{v_{out}}{v_{in}} &= 1+\frac{R_2}{R_1} \end{aligned} \tag{5.11}$$

This result is accurate, as anyone who has tested such an amplifier in the lab may attest.

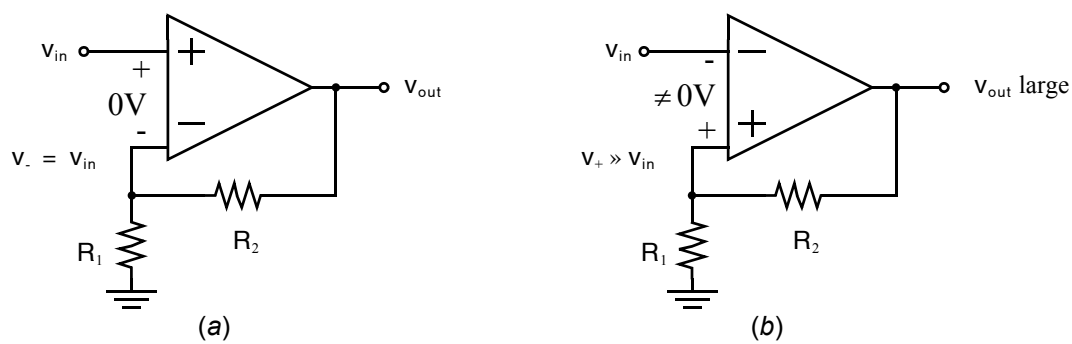


Fig. 5.3 (*a*) A noninverting amplifier. (*b*) A noninverting amplifier with the opamp input terminals reversed is unstable.

One might be tempted to apply the same analysis to the amplifier in Fig. 5.3(*b*). Doing so would yield the same result (5.11) since the only difference between the circuits is the polarity of the opamp inputs which was nowhere considered in the analysis above. However, assuming that the opamp's differential input is zero presumes stability of the feedback loop. The feedback loop in Fig. 5.3(*b*) is in fact unstable. In practice, the voltage at v_- will increase above v_{in}, and the opamp's large gain will amplify the difference producing an output voltage v_{out} close to the amplifier's supply voltage.

Example 5.3 is an extreme case where the loop's feedback is positive instead of negative. However, instability can arise more subtly, when at high frequencies the amplifier introduces a phase shift that causes negative feedback to become positive. In this section, general methods are presented to check the stability of analog circuits with feedback. For stable circuits the same methods provide insight into important aspects of the closed-loop amplifier's frequency response.

5.2.1 Stability Criteria

The methods applied to study the stability of analog circuits with feedback differ somewhat from those employed in courses and texts on introductory control theory. Control theory is often concerned with the problem of stabilizing, by application of feedback, a system that is otherwise unstable.[2] By contrast, we will apply feedback to amplifiers A that are already stable on their own. We use feedback to improve the amplifier's accuracy, bandwidth, and linearity, as described in Section 5.1. Unfortunately, if care is not taken, feedback can make the previously-stable amplifier go unstable, as in Example 5.3.

Stability can be tested by checking that all poles of the closed loop system,

$$A_{CL}(s) = \frac{A(s)}{1 + A(s)\beta} \tag{5.12}$$

are in the left-half plane. However, this method is not often used because obtaining $A(s)$ with accuracy is difficult. Fortunately, stability can be checked with knowledge of the magnitude and phase of the frequency response $L(\omega)$, which are readily obtained by circuit simulation.

In all situations of interest to us, the amplifier's magnitude response is large at low frequencies, $|A(\omega)| \gg 1/\beta$, and decreases at high frequencies. The same poles that cause the magnitude response to roll off also contribute negative phase shift at higher frequencies. Furthermore, in accordance with the discussion in Section we shall assume that the feedback circuit has a constant and well-controlled gain, $\beta \leq 1$. The open-loop magnitude response is, therefore, simply a scaled version of the amplifier's magnitude response,

$$|L(\omega)| = |A(\omega)|\beta \tag{5.13}$$

and the open-loop phase response is equal to that of the amplifier,

$$\angle L(\omega) = \angle A(\omega) + \angle\beta = \angle A(\omega) + 0° = \angle A(\omega) \tag{5.14}$$

The open-loop Bode plot sketched in Fig. 5.4(*a*) is typical of all the feedback circuits we will consider. Of particular interest is the frequency at which the open-loop magnitude response is unity,

$$|L(\omega_t)| = 1 \tag{5.15}$$

The solution of (5.15) is the loop's "unity-gain frequency", ω_t.[3]

---

2. The system to be stabilized is referred to as "the plant."

3. Barring any strange ripples in the loop's magnitude response (which should certainly be avoided by design) equation (5.15) has only one (positive) solution and ω_t is unique.

An alternative graphical representation of the open-loop magnitude and phase response is shown in Fig. 5.4(*b*), which is a polar plot of $L(\omega) = A(\omega)\beta$ for all frequencies, $-\infty < \omega < \infty$. At low frequencies where $L(\omega)$ has large magnitude and near-zero phase shift, the polar plot will be far from the origin along the positive real axis. As ω increases to large positive values, the plot spirals clockwise inwards to the origin. For negative values of ω, the spiral is mirrored in the real axis, spiraling counterclockwise from the point at $L(0)$. Given a stable amplifier A, the closed-loop feedback system will also be stable if and only if this polar plot does not encircle the point -1.[4] Such will be the case so long as the polar plot intersects the unit circle below the real axis for positive frequencies, as is the case in Fig. 5.4.

By definition, the polar plot's intersection with the unit circle occurs at the loop's unity-gain frequency, ω_t. Hence, an equivalent stability criterion is

$$\angle L(\omega_t) > -180° \tag{5.16}$$

Key Point: *The stability of a feedback amplifier may be checked by looking at a Bode plot of the loop gain, $L(\omega)$; so long at the phase response is greater than –180 degres at the unity gain frequency, the feedback system is stable.*

In summary, a feedback amplifier is stable so long as (5.16) is satisfied with ω_t defied by (5.15). This criteria is written completely in terms of the open-loop magnitude and phase response. Hence, it can be checked simply by examining a Bode plot of $L(\omega)$. There is little need for the polar plots and they will henceforth be avoided. Furthermore, using the procedure in Section 5.4.1, the Bode plot may be readily obtained by computer simulation.

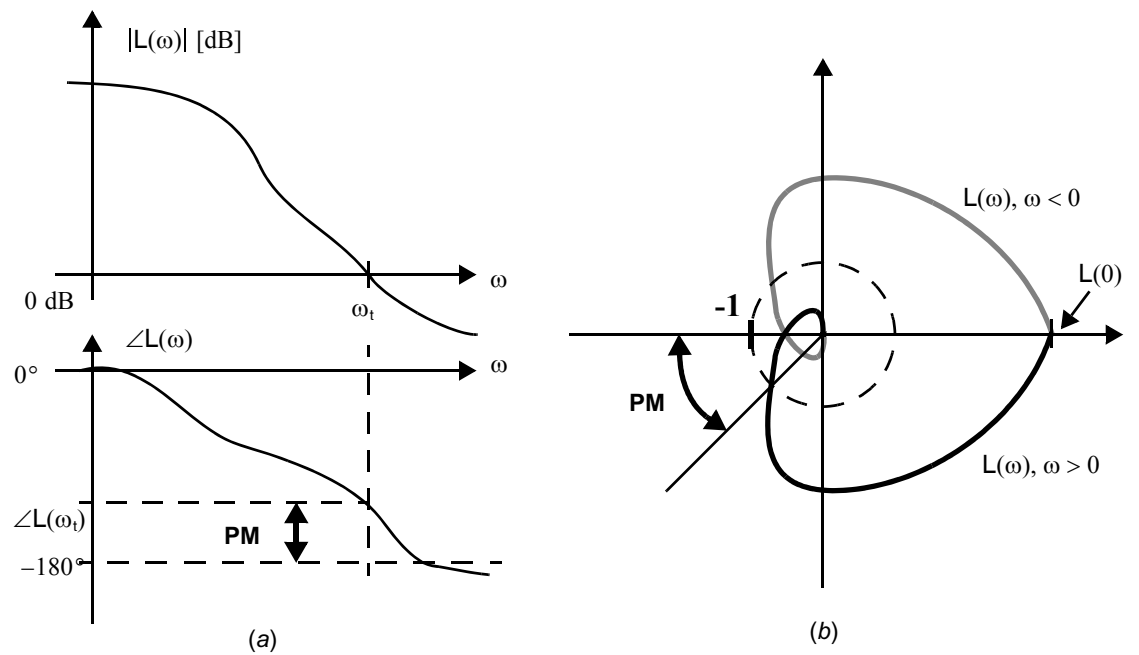


Fig. 5.4 The relationship between (*a*) a Bode plot and (*b*) a polar plot of $L(\omega)$ for a stable feedback circuit.

4. This is special case of the well-known "Nyquist stability criterion," and the polar plot of $L(\omega)$ is a "Nyquist plot." For a general proof, the reader is referred to introductory control theory texts.

EXAMPLE 5.4

Sketch a polar plot of loop gain for the positive-feedback circuit in Fig. 5.3(*b*).

Solution

Since the feedback is positive, when cast into our negative feedback model of Fig. 5.1 the loop gain $\mathsf{L} = \mathsf{A}\beta$ is negative at low frequencies where the opamp gain is very high. Hence, the plot begins on the negative real axis with large magnitude. From there, the magnitude response of the loop will tend to zero at high frequencies, so the polar plot spirals into the origin, as shown in Fig. 5.5. Clearly, it must encircle the point –1 so the loop is unstable.

5.2.2 Phase Margin

As the name suggests, phase margin provides a quantitative measure of how close a given feedback system is to instability. However, phase margin also provides the designer with important information about the closed-loop response of a feedback amplifier.

Phase margin is defined for stable feedback systems as the additional phase shift that would be required at the unity-gain frequency to cause instability. The definition is illustrated graphically in Fig. 5.4. Analytically, phase margin is

$$\mathsf{PM} \equiv \angle \mathsf{L}(\omega_t) + 180^\circ \tag{5.17}$$

It is commonplace to establish some minimum phase margin as a basic specification in the design of a feedback amplifier. A typical misconception amongst students of analog design is that such specifications are intended to safeguard against instability in the presence of variations in circuit parameter values. However, we generally demand far more phase margin than is required simply to ensure the system is robustly stable. Rather, large phase margins of between 45° and 90° are typically required because systems with smaller phase margins will exhibit undesirable dynamic behavior.

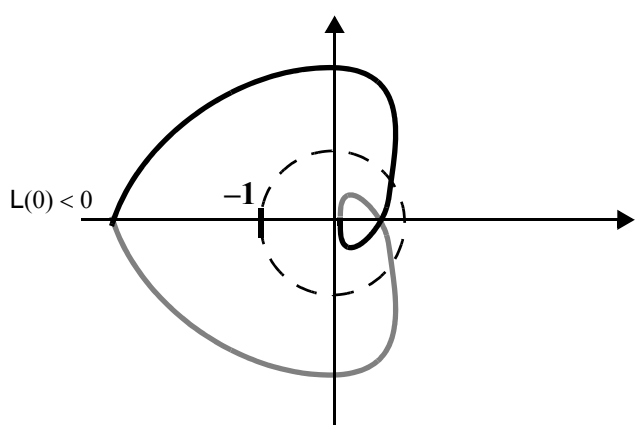


Fig. 5.5 Polar plot of the loop gain for the case of positive feedback. Since the L is large and negative at low frequencies, the plot must certainly encircle the point –1 and the loop is therefore unstable.

EXAMPLE 5.5

Find an expression for the closed-loop gain of a feedback amplifier at the unity gain frequency, ω_t. Assume the system has a phase margin θ.

Solution

At the unity gain frequency, combining (5.15) and (5.17) reveals

$$L(\omega_t) = -e^{j\theta} \tag{5.18}$$

$$\Rightarrow A(\omega_t) = \frac{L(\omega_t)}{\beta} = -\frac{e^{j\theta}}{\beta} \tag{5.19}$$

Substituting (5.18) and (5.19) into (5.6) yields,

$$A_{CL}(\omega_t) = \frac{-e^{j\theta}/\beta}{1 - e^{j\theta}} \tag{5.20}$$

Taking the absolute value and recognizing that $|1 - e^{j\theta}| = \sqrt{2(1 - \cos\theta)}$,

$$|A_{CL}(\omega_t)| = \frac{1}{\beta\sqrt{2(1 - \cos\theta)}} \tag{5.21}$$

Notice that for $\theta < 60°$, $|A_{CL}(\omega_t)| > 1/\beta$. But, it has already been shown in (5.9) that the closed-loop gain of such an amplifier at low-frequencies is

$$A_{CL} \cong 1/\beta$$

Hence, with a phase margin less than $60°$, the closed-loop gain actually increases as one moves from low frequencies up to ω_t! Of course, at even higher frequencies the gain eventually decreases and we have $A_{CL}(\omega) \cong A(\omega) \ll (1/\beta)$.

Key Point: *It is not enough that an amplifier simply be reliably stable—it must also reliably have large phase margin. To avoid undesirable behavior in the frequency and step responses of the closed-loop amplifier, it is common to require phase margins of 45 to 90 degrees.*

Example 5.5 reveals that the sketch of closed-loop frequency response in Fig. 5.2 is accurate only for phase margins greater than $60°$. Systems with phase margins less than $60°$, but still stable, have a closed loop-frequency response as in Fig. 5.6, where peaking is observed around the loop's unity gain frequency. Since we wish the gain of the closed-loop amplifier to be constant around $1/\beta$, such peaking is undesirable and, as we shall see in later sections, is generally accompanied by overshoot in the amplifier's transient step response. Hence, it is not enough that an amplifier simply be reliably stable—it must also reliably have large phase margin.

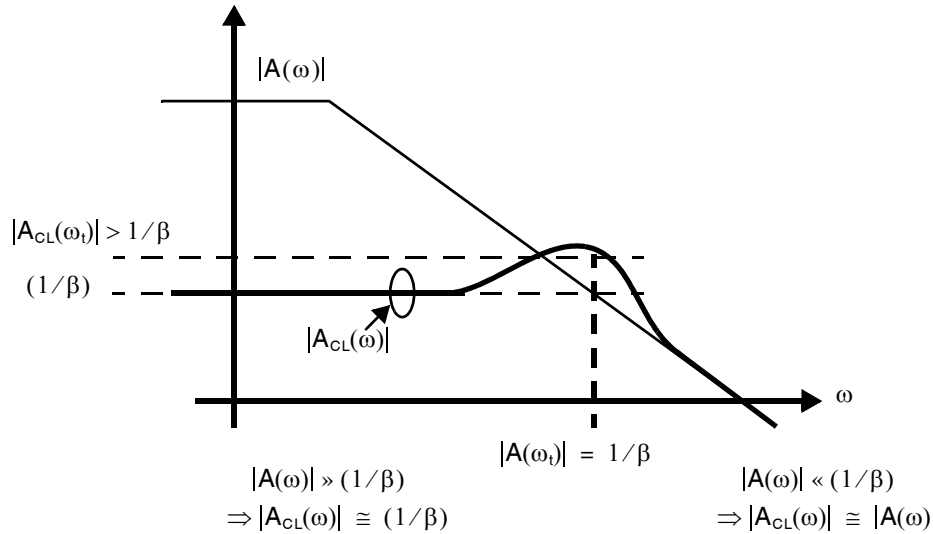


Fig. 5.6 The closed-loop frequency response of a feedback amplifier with a phase margin less than 60 degrees.

5.3 FIRST- AND SECOND-ORDER FEEDBACK SYSTEMS

Although the frequency response of a practical amplifier may involve many poles and zeros, it is often sufficient to model it with a simple first- or second-order transfer function. It is most important to accurately understand the amplifier's response at frequencies $\omega \lesssim \omega_t$ which determines the amplifier's stability and in-band performance. Hence, any poles and zeros at frequencies $\omega \gg \omega_t$ may be neglected. A typical situation is depicted in Fig. 5.7 where one pole is dominant. A simple first-order model is sufficient to predict the behavior of feedback amplifiers when the unity-gain frequency is well below the second- and higher-order pole- and zero-frequencies of the amplifier, $\omega_t \ll \omega_{p2}, \omega_{p3}, |\omega_{z1}|, \ldots$ The second-order model is preferred when ω_t is closer to $\omega_{p2}, \omega_{p3}, \omega_{z1}, \ldots$ For cases where ω_t is at a frequency higher than $\omega_{p2}, \omega_{p3}, \omega_{z1}, \ldots$, even the second-order model has insufficient accuracy. However, these cases will have very little phase margin and they are, therefore, of little practical interest. This section focuses on simple first- and second-order models providing insight into the behavior of practical feedback amplifiers, which often defy rigorous analytical treatment.

5.3.1 First-Order Feedback Systems

A simple first-order model for the transfer function of a dominant-pole feedback amplifier, $AL(s)$, with dc gain L_0 and dominant pole frequency ω_{p1} is given by

$$L(s) = \frac{L_0}{(1 + s/\omega_{p1})} \tag{5.22}$$

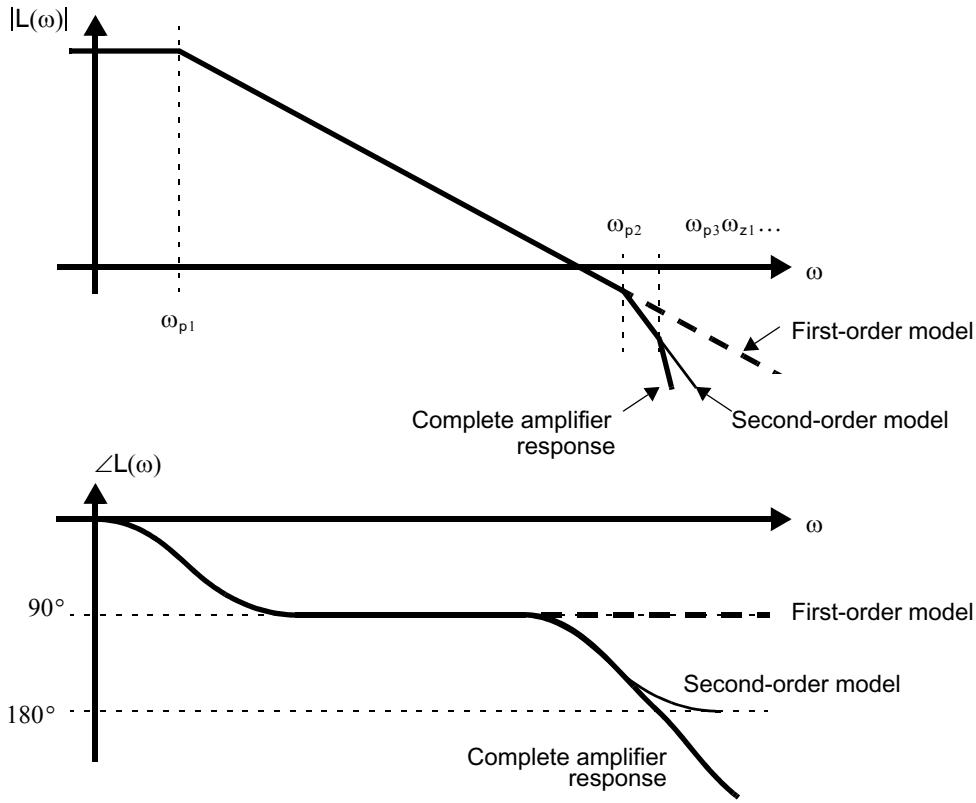


Fig. 5.7 Typical feedback amplifier open-loop frequency response and the corresponding first- and second-order models.

Since there is only one pole contributing phase shift, the phase response of a first-order system never drops below –90°. L,

$$\angle L(\omega) > -90^\circ \tag{5.23}$$

Key Point: *A first-order feedback amplifier is unconditionally stable with approximately 90 degrees phase margin.*

Combining (5.23) with the stability criteria in (5.16) and the definition of phase margin in (5.17), it is clear that a first-order feedback system is stable with at least 90° phase margin. It is evident from the Bode plot in Fig. 5.8(*a*) that for large dc open-loop gain, L_0, the phase margin will in fact be very close to 90°. Furthermore, it is impossible for the polar plot of $L(\omega)$ for a negative feedback system with only 1 pole to encircle the point –1, Fig. 5.8(*b*). Hence, first-order feedback systems are stable with 90° phase margin *for any dc gain, or pole frequency.* This is unlike second- and higher-order loops where phase margin critically depends upon the feedback factor, β, dc gain, and pole frequencies,

Recall the unity-gain frequency ω_t, is the frequency at which $|L(\omega_t)| = 1$. We then have the following approximation, assuming $\omega_t >> \omega_{p1}$:

$$|L(\omega_t)| = 1 \cong \frac{L_0}{\omega_t / \omega_{p1}} \tag{5.24}$$

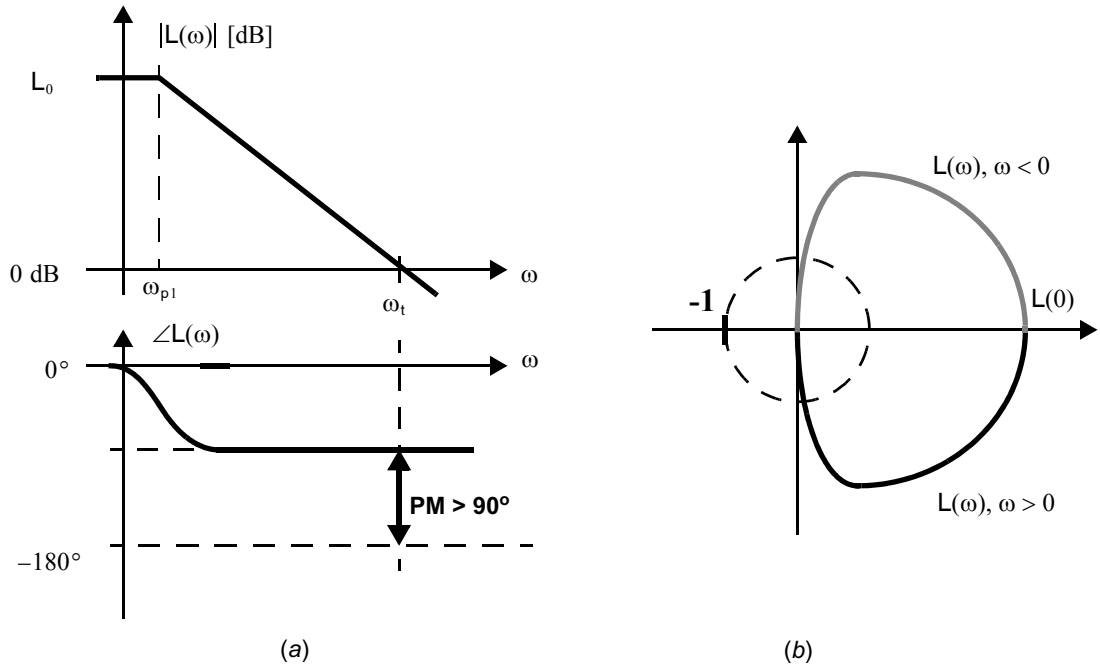


Fig. 5.8 Graphical illustration of the unconditional stability and 90° phase margin of first-order feedback systems: (*a*) Bode plot and (*b*) polar plot of open-loop frequency response, $L(\omega)$.

Thus, we have the following important relationship for this first-order model:

$$\omega_t \cong L_0\omega_{p1} \tag{5.25}$$

Substituting (5.25) into (5.22) for the case in which $\omega \gg \omega_{p1}$, we have at midband frequencies

$$L(s) \cong \frac{\omega_t}{s} \tag{5.26}$$

The relationship (5.26) is a simplified first-order model that assumes infinite dc gain, and is therefore not accurate at low frequencies. However, it can be used to analyze the system at midband frequencies, $\omega_{p1} \ll \omega \ll \omega_t$, including estimating the closed-loop circuit's –3dB frequency and settling time.

At midband frequencies, the transfer function of the closed-loop amplifier may be found by substituting (5.26) into (5.8), resulting in

$$A_{CL}(s) \cong \frac{1}{\beta}\frac{1}{(1 + s/\omega_t)} \tag{5.27}$$

As expected, the closed-loop amplifier has a closed-loop gain at low frequencies approximately equal to $1/\beta$. The –3dB frequency of the closed-loop amplifier is given by

$$\omega_{-3\,dB} \cong \omega_t \tag{5.28}$$

Furthermore, from (5.27) we recognize that the closed-loop amplifier is a first-order system with a time-constant, τ, given by

$$\tau = \frac{1}{\omega_{-3\,dB}} = \frac{1}{\omega_t} \tag{5.29}$$

The settling-time performance of integrated amplifiers is often an important design parameter. For example, in switched-capacitor circuits, the charge from one or more capacitors must be mostly transferred to a feedback capacitor within about half a clock period. This charge transfer is closely related to the opamp's step response. As a result, the *settling time* is defined to be the time it takes for an opamp to reach a specified percentage of its final value when a step input is applied.

Recall that the step response of any first-order circuit is given by

$$v_{out}(t) = V_{step}(1 - e^{-t/\tau}) \tag{5.30}$$

Here, V_{step} is the size of the voltage step. With this exponential relationship, the time required for the first-order closed-loop circuit to settle with a specified accuracy can be found. For example, if 1% accuracy is required, then one must allow $e^{-t/\tau}$ to reach 0.01, which is achieved at a time of $t = 4.6\tau$. For settling to within 0.1% accuracy, the settling time needed becomes approximately 7τ. Also, note that just after the step input, the slope of the output will be at its maximum, given by

$$\left.\frac{d}{dt}v_{out}(t)\right|_{t=0} = \frac{V_{step}}{\tau} \tag{5.31}$$

Key Point: *The closed-loop bandwidth of a feedback amplifier is ω_t, independent of the amplifier's 3-dB bandwidth, ω_{p1}. The settling time is $1/\omega_t$.*

Since a real amplifier will have a finite current available to charge the parasitic capacitances at v_{out}, in many practical cases for large steps V_{step} the amplifier will be unable to change the output voltage at the rate predicted by (5.31). In such a case, the amplifier is no longer behaving linearly and the model's predictions break down.

EXAMPLE 5.6

One phase of a switched-capacitor circuit is shown in Fig. 5.9, where the input signal can be modelled as a voltage step and 0.1% accuracy is needed in 0.1 μs. Assuming linear settling, find the required unity-gain frequency in terms of the capacitance values, C_1 and C_2. For $C_2 = 10C_1$, what is the necessary unity-gain frequency of the opamp? What unity-gain frequency is needed in the case $C_2 = 0.2C_1$?

Solution

We first note that a capacitive feedback network is used rather than a resistive one. A difficulty with this network in a nonswitched circuit is that no bias current can flow into the negative opamp input terminal. However, in a switched circuit using a CMOS opamp, the shown configuration occurs only for a short time and does not cause any problems. Second, assume the opamp has a dominant pole response with unity gain frequency ω_{ta} (different from the unity gain frequency of $L(s)$, defined as ω_t) and large dc gain. Hence, its frequency response at midband

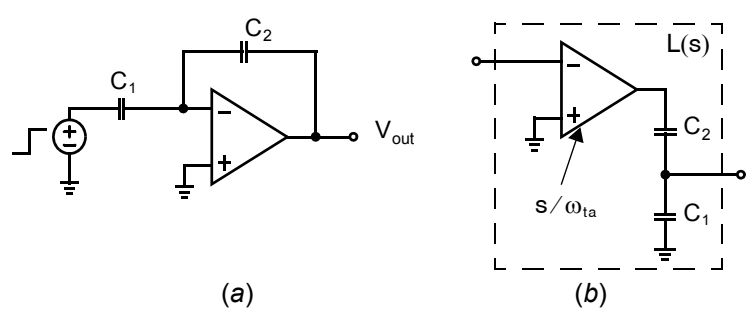


Fig. 5.9 (*a*) One phase of a switched-capacitor circuit. (*b*) Open-loop feedback circuit illustrating $L(s)$

frequencies is well approximated by s/ω_{ta}. The open-loop response, $L(s)$, is given by this opamp response driving the capacitive feedback network, as shown in Fig. 5.9(*b*).

$$L(s) \cong \left(\frac{s}{\omega_{ta}}\right)\left(\frac{C_2}{C_1 + C_2}\right) \tag{5.32}$$

The unity gain frequency of $L(s)$ is found by substituting (5.32) into (5.15) and solving for $\omega_t = \omega_{ta}C_2/(C_1 + C_2)$. For 7τ settling within 0.1 μs, we see that τ must be less than 14.2 ns. Since $\tau = 1/\omega_t$,

$$\omega_{ta} \geq \left(\frac{C_1 + C_2}{C_2}\right)\left(\frac{1}{14.2\text{ ns}}\right) \tag{5.33}$$

For the case in which $C_2 = 10C_1$, a unity-gain frequency of $2\pi \cdot 12.3$ MHz is needed, whereas in the case of $C_2 = 0.2C_1$, ω_{ta} should be larger than $2\pi \cdot 66.8$ MHz.

The unconditional stability, predictable bandwidth, and predictable settling time of first-order systems are strong incentives to ensure feedback circuits have a dominant pole and, therefore, may be approximated by the single-pole model in (5.22).

5.3.2 Second-Order Feedback Systems

Although first-order feedback amplifiers exhibit several desirable characteristics, practical feedback amplifiers generally have more than one pole. Recall that the main objective in the design of the amplifier A is that it must have large low-frequency gain. Unfortunately, the maximum gain achievable with a single transistor, given by its intrinsic gain $A_i = g_m r_{ds}$, is usually insufficient. Hence, multiple transistors must be combined resulting in a more complex circuit with several nodes and, inevitably, multiple poles.

A general second-order feedback loop transfer function is,

$$L(s) = \frac{L_0}{(1 + s/\omega_{p1})(1 + s/\omega_{eq})} \tag{5.34}$$

It is assumed that $\omega_{eq} > \omega_{p1}$. It is clear from the Bode plot of (5.34) in Fig. 5.7 that it will be unconditionally stable since a phase shift of $-180°$ is never attained. However, the phase margin may approach $0°$ depending on L_0, ω_{p1}, and ω_{eq}.

Key Point: *All-pole second-order feedback systems are unconditionally stable with a phase margin of between 90 and 0 degrees.*

At frequencies much greater than the dominant pole frequency, $\omega >> \omega_{p1}$, we see that $1 + j\omega/\omega_{p1} \cong j\omega/\omega_{p1}$, and so (5.34) can be accurately approximated by

$$L(s) \cong \frac{\omega_t}{s(1 + s/\omega_{eq})} \tag{5.35}$$

Note that this approximation result is especially valid at the unity-gain frequency of the loop ω_t (which we are presently interested in) which is almost certainly much greater than ω_{p1} as long as $L_0 \gg 1$.

Since $L(s) = \beta A(s)$ where β is a scalar constant, it is clear that the poles are contributed by the forward amplifier, A(s). Using (5.34) we can define $A_0 = L_0/\beta$ which is the dc gain of A(s) and $\omega_{ta} = A_0\omega_{p1}$ as the approximate unity-gain frequency of A(s) under a dominant pole approximation. Hence, at frequencies around ω_t, the loop gain, $L(s)$, is given by

$$L(s) = \beta A(s) = \frac{\beta\omega_{ta}}{s(1 + s/\omega_{eq})} \tag{5.36}$$

The unity-gain frequency, ω_t, can now be found by setting the magnitude of (5.36) equal to unity after substituting $s = j\omega_t$. Once this is done and the equation is rearranged, one can write,

$$\beta\frac{\omega_{ta}}{\omega_{eq}} = \frac{\omega_t}{\omega_{eq}}\sqrt{1+\left(\frac{\omega_t}{\omega_{eq}}\right)^2} \tag{5.37}$$

For the dominant pole special case in which the unity-gain frequency is much less than the equivalent nondominant pole frequency (i.e., $\omega_t \ll \omega_{eq}$), (5.37) may be simplified to

$$\omega_{ta} = \frac{\omega_t}{\beta}\sqrt{1+\left(\frac{\omega_t}{\omega_{eq}}\right)^2} \cong \frac{\omega_t}{\beta} \tag{5.38}$$

From (5.36), the phase shift, $\angle L(\omega)$, is found.

$$\angle L(\omega) = -90° - \tan^{-1}(\omega/\omega_{eq}) \tag{5.39}$$

This equation implies that at the unity-gain frequency, $\omega = \omega_t$, we have

$$PM = \angle L(\omega_t) - (-180°) = 90° - \tan^{-1}(\omega_t/\omega_{eq}) \tag{5.40}$$

and, therefore,

$$\begin{aligned} \omega_t/\omega_{eq} &= \tan(90° - PM) \\ \Rightarrow \omega_t &= \tan(90° - PM)\omega_{eq} \end{aligned} \tag{5.41}$$

Key Point: *The process of designing or modifying a feedback amplifier to have some targeted phase margin is called compensation.*

Equation (5.41) gives the required ω_t/ω_{eq} for a specified phase margin. The process of designing or modifying a feedback amplifier to have the ratio ω_t/ω_{eq} required by (5.41) for some targeted phase margin is called *compensation*.

EXAMPLE 5.7

A closed-loop amplifier is to have a $75°$ phase margin for $\beta = 1$. What is the required unity-gain frequency f_t if $f_{eq} = \omega_{eq}/(2\pi) = 50\,\text{MHz}$? What is the required f_{ta}?

Solution

Using (5.40), we have $\omega_t = 0.268\omega_{eq}$, which implies that the loop-gain unity-gain frequency is given by $f_t = \omega_t/2\pi = 13.4$ MHz. Using (5.38), we also have, for $\beta = 1$, $\omega_{ta} = \omega_t\sqrt{1+0.268^2} = 1.035\omega_t$, which implies that $f_{ta} = \omega_{ta}/2\pi = 13.9$ MHz.

The closed-loop response of a second-order feedback amplifier may be found by substituting (5.34) into (5.8) and rearranging.

$$A_{CL}(s) = \frac{A_{CL0}}{1 + \dfrac{s(1/\omega_{p1} + 1/\omega_{eq})}{1+L_0} + \dfrac{s^2}{(1+L_0)(\omega_{p1}\omega_{eq})}} \tag{5.42}$$

where

$$A_{CL0} \equiv \frac{L_0/\beta}{1+L_0} \cong \frac{1}{\beta} \tag{5.43}$$

Table 5.1 The relationship between PM, ω_t/ω_{eq}, Q factor, and percentage overshoot

PM (Phase margin)	ω_t/ω_{eq}	Q factor	Percentage overshoot for a step input
55°	0.700	0.925	13.3%
60°	0.580	0.817	8.7%
65°	0.470	0.717	4.7%
70°	0.360	0.622	1.4%
75°	0.270	0.527	0.008%
80°	0.175	0.421	-
85°	0.087	0.296	-

Thus, the closed-loop response is also second-order. The general equation for a second-order all-pole transfer function is

$$H(s) = \frac{K\omega_0^2}{s^2 + \left(\frac{\omega_0}{Q}\right)s + \omega_0^2} = \frac{K}{1 + \frac{s}{\omega_0 Q} + \frac{s^2}{\omega_0^2}} \tag{5.44}$$

Recall from the preceding chapter that ω_0 is called the resonant frequency and parameter Q is called the Q factor[5] [Sedra, 1991]. If $Q = \sqrt{1/2} \cong 0.707$, then the magnitude response will have a –3-dB frequency equal to ω_0. Lower Q-factors ($Q < \sqrt{1/2}$) result in lower bandwidth whereas higher Q-factors ($Q > \sqrt{1/2}$) result in peaking in the frequency response. Furthermore, recall that if $Q \le 0.5$ both poles of $A_{CL}(s)$ are real and there is no overshoot in the feedback amplifier's step response. In the case where $Q > 0.5$, the percentage overshoot of the step response is given by

$$\% \text{ overshoot} = 100e^{\frac{-\pi}{\sqrt{4Q^2-1}}} \tag{5.45}$$

Equating (5.42) with (5.44) and solving for ω_0 and Q results in

$$\omega_0 = \sqrt[t]{(1 + L_0)(\omega_{p1}\omega_{eq})} \cong \sqrt{\omega_i \omega_{eq}} \tag{5.46}$$

and

$$Q = \frac{\sqrt{(1 + L_0)/\omega_{p1}\omega_{eq}}}{1/\omega_{p1} + 1/\omega_{eq}} \cong \sqrt{\frac{L_0\omega_{p1}}{\omega_{eq}}} = \sqrt{\frac{\beta\omega_{ta}}{\omega_{eq}}} \tag{5.47}$$

where the approximation on Q is valid since $L_0 >> 1$ and $\omega_{p1} << \omega_{eq}$.

It is now possible to relate a specified phase margin to the Q factor of the resulting closed-loop amplifier. Equation (5.41) can be used to find ω_t/ω_{eq}. This result can be substituted into (5.37) to find $\beta(\omega_{ta}/\omega_{eq})$, which can then be substituted into (5.47) to find the equivalent Q factor. Finally, (5.45) can be used to find the corresponding percentage overshoot for a step input. This procedure gives us the information in Table 5.1.

Table 5.1 leads to some interesting observations. First, a frequency response with $Q \cong \sqrt{1/2}$ roughly corresponds to a phase margin of 65°. Therefore, a common specification is to design for a phase margin of at least $65°$ in the presence of both process and temperature changes as this ensures no peaking in the closed-loop amplifier's frequency response. If one wants to ensure that there is no overshoot for a step input, then the phase margin should be at least 75°, again, given both process and temperature variations. Hence, a nominal phase margin of $80°$ to $85°$ is often targeted to account for these variations.

5. The Q factor is 1/2 times the inverse of the damping factor. The damping factor is an alternative method of indicating the pole locations in second-order transfer functions.

Finally, it is worth mentioning that when an amplifier is designed for use with many different feedback networks, the worst-case phase margin occurs for the maximum value of β which corresponds to the smallest closed-loop gain. Thus, the amplifier is designed to have sufficient phase margin for the maximum value of β. It is then guaranteed to have even greater phase margin for all other β, although in these cases it will be sub-optimally compensated and will be slower than necessary.

5.3.3 Higher-Order Feedback Systems

High-gain amplifiers will often include more than two poles, and perhaps some zeroes. A general transfer function for such an amplifier is

$$A(s) = A_0 \frac{\left(1+\frac{s}{\omega_{z1}}\right)\left(1+\frac{s}{\omega_{z2}}\right)\ldots\left(1+\frac{s}{\omega_{zm}}\right)}{\left(1+\frac{s}{\omega_{p1}}\right)\left(1+\frac{s}{\omega_{p2}}\right)\ldots\left(1+\frac{s}{\omega_{pn}}\right)} \tag{5.48}$$

Feedback systems based on such amplifiers are not necessarily stable since, at some frequency, they may attain a phase shift beyond 180°.

It is important to accurately model the open-loop transfer function at frequencies up to and around the loop's unity-gain frequency, ω_t. This can be done with the first-order model (5.22) only when the second and higher-order poles and all zeros are at frequencies much higher than ω_t,

$$\omega_t \ll \omega_{p2}, \ldots, \omega_{pn}, |\omega_{z1}|, \ldots, |\omega_{zm}| \tag{5.49}$$

Even when it is possible to design an amplifier that satisfies (5.49), doing so is often overly-conservative resulting in significant over-design and a sub-optimal circuit. Instead, by accounting for all poles and zeros the specified phase margin can be achieved with a higher unity-gain frequency, ω_t, which as we saw in Section 5.1.3 will result in a higher closed-loop bandwidth.

Key Point: *Higher-order feedback amplifiers may be modeled using a second-order transfer function so long at they are not approaching instability. The second pole is chosen as the frequency at which a phase response of –135 degrees is attained.*

When all poles and zeros are on the real axis, they can be modelled reasonably well by just two poles. Specifically, we include in our model the first dominant-pole frequency, ω_{p1}, and a second pole frequency that models all higher-frequency poles and zeros, ω_{eq}. Given a set of real-axis poles, ω_{pi}, and zeros, ω_{zi}, a good choice for ω_{eq} is given by

$$\frac{1}{\omega_{eq}} \cong \sum_{i=2}^{n} \frac{1}{\omega_{pi}} - \sum_{i=1}^{m} \frac{1}{\omega_{zi}} \tag{5.50}$$

It should be noted here that the approximation in (5.50) is different than that given in [Sedra, 1991] since we are mostly interested in the phase shifts due to higher-frequency poles and zeros, rather than the attenuation. In practice, ω_{eq} is found from simulation as the frequency at which the open-loop transfer function, $L(s)$, has a –135° phase shift (–90° due to the dominant pole and another –45° due to the combined effect of higher-frequency poles and zeros).

5.4 COMMON FEEDBACK AMPLIFIERS

We will now apply our general analysis of feedback systems to small-signal circuits. In order to do so, we must cast the circuit of interest into the model of Fig. 5.1. Unfortunately, this is often easier said than done. There may be ambiguity in identifying whether the inputs and/or outputs of a particular circuit are voltages or currents. Furthermore, the forward amplifier, A, and feedback network, β, are difficult to discern. Some components in a circuit are simultaneously involved in both the forward amplifier and the feedback network.

In particular, it is difficult to find A. For example, consider the inverting amplifier configuration in Fig. 5.10 where the opamp has a finite output impedance, Z_o. It is tempting to conclude that A is simply the voltage gain of the opamp, but this is not the case. Equation (5.5) and Fig. 5.1 show that when $A = 0$ the closed-loop amplifier output must be zero. However, even if the gain of the opamp were zero, a finite voltage would appear at v_{out} due to division of the source voltage, v_s, by impedances in the feedback network (Z_1 and Z_2) and at the output node (Z_o and Z_L). In fact, an accurate analysis for A is quite complex even in this every-day example. It depends not only on the opamp but also the feedback components Z_1 and Z_2, and the load, Z_L. Computer simulation is of little help since it is not obvious exactly what circuit must be simulated to find A, the forward gain without feedback.

On the other hand, the value of β is usually relatively obvious if one remembers that the circuit's desired gain is simply $1/\beta$. For example, the gain we expect from the inverting opamp configuration above is well-known to be $-(Z_2/Z_1)$. Hence, we may assume that $\beta = -(Z_1/Z_2)$. More generally, the value of β is the inverse of the closed-loop gain we expect under ideal conditions of very large loop gain. In opamp circuits, this gain will result from analyzing the circuit assuming infinite opamp gain. Although this choice of β does not necessarily identify the gain between any two particular voltages and/or currents in the circuit, this need not trouble us—it clearly is in keeping with our immediate goal of casting the circuit into the ideal feedback model, and more importantly allows us to find the properties of a feedback circuit that are of most interest to us such as gain error, bandwidth, phase margin, and settling time.

Whereas there may be some ambiguity in defining A, the loop gain L is unique. This is satisfying since stability of the loop depends only on $L(s)$, and physically a feedback circuit will be either stable of unstable, regardless of how we choose to analyze it.

If L is known and β is taken by definition to be the reciprocal of the desired closed-loop gain, rather than trying to perform a rigorous analysis for A we can infer it from $A = L/\beta$. Therefore, our analysis of feedback circuits will be based on a knowledge of β, ours either by design or by inspection of a relatively simply circuit, and from a knowledge of L which can be estimated using a systematic procedure described in Section 5.4.1.

Generalizing the expression for closed-loop gain in (5.8), the closed-loop frequency response of a feedback circuit can be expressed without reference to $A(s)$.

$$A_{CL}(s) = \frac{1}{\beta} \cdot \frac{L(s)}{1 + L(s)} \tag{5.51}$$

In (5.51) the term $L(s)/(1 + L(s))$ is the closed-loop amplifier gain normalized to the desired gain, $1/\beta$. Many circuit attributes of interest to us may be thought of in terms of this normalized frequency response which depends only upon $L(s)$. For example, the bandwidth, the poles and the zeros of the closed-loop amplifier are those of $L(s)/(1 + L(s))$. The *gain error* at dc, defined as the fractional error between the desired gain, $1/\beta$, and the actual gain, $A_{CL0} = A_{CL}(0)$, is simply

$$\text{Gain error} = 1 - \frac{A_{CL0}}{1/\beta} = 1 - \frac{L_0}{1 + L_0} \tag{5.52}$$

which will clearly assume a value very close to zero so long as the loop gain at dc, $L_0 = L(0)$, is large.

> ***Key Point:*** *Feedback systems may be analyzed in terms of their desired gain, β, and loop gain, $L(\omega)$, from which the closed-loop amplifier's gain, bandwidth, phase margin, and frequency response may all be inferred.*

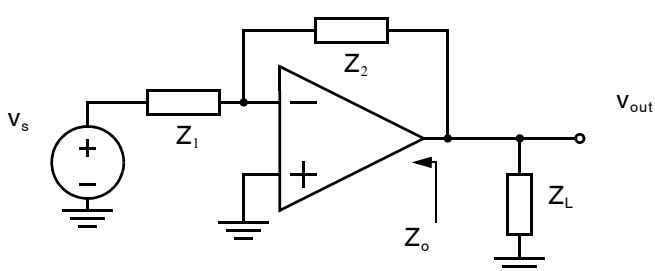


Fig. 5.10 An inverting opamp configuration.

5.4.1 Obtaining the Loop Gain, L(s)

To find the loop gain of a feedback circuit, we break the loop, insert a test signal (either voltage or current) and see what signal the loop returns to the other side of the break. The procedure is illustrated in Fig. 5.11. Specifically, for any (small-signal) feedback circuit,

1. Set any independent sources in the circuit equal to zero, including the input source. Zeroing a voltage source means replacing it will a short circuit. (A short circuit is an ideal voltage source of 0 V.) Zeroing a current source means replacing it with an open circuit. (An open circuit is an ideal current source with a value of 0 A.)
2. Break the loop. Find the impedance at the break point, Z_t, and terminate the loop with this impedance as shown in Fig. 5.11.
3. Insert a test signal (either voltage, v_t, or current, i_t) into the loop at the break and find the returned signal: either the voltage across Z_t, v_r, or the current through Z_t, i_r. The loop gain is then

$$L = -\frac{v_r}{v_t} \text{ (voltage test signals) or } -\frac{i_r}{i_t} \text{ (current test signals).} \tag{5.53}$$

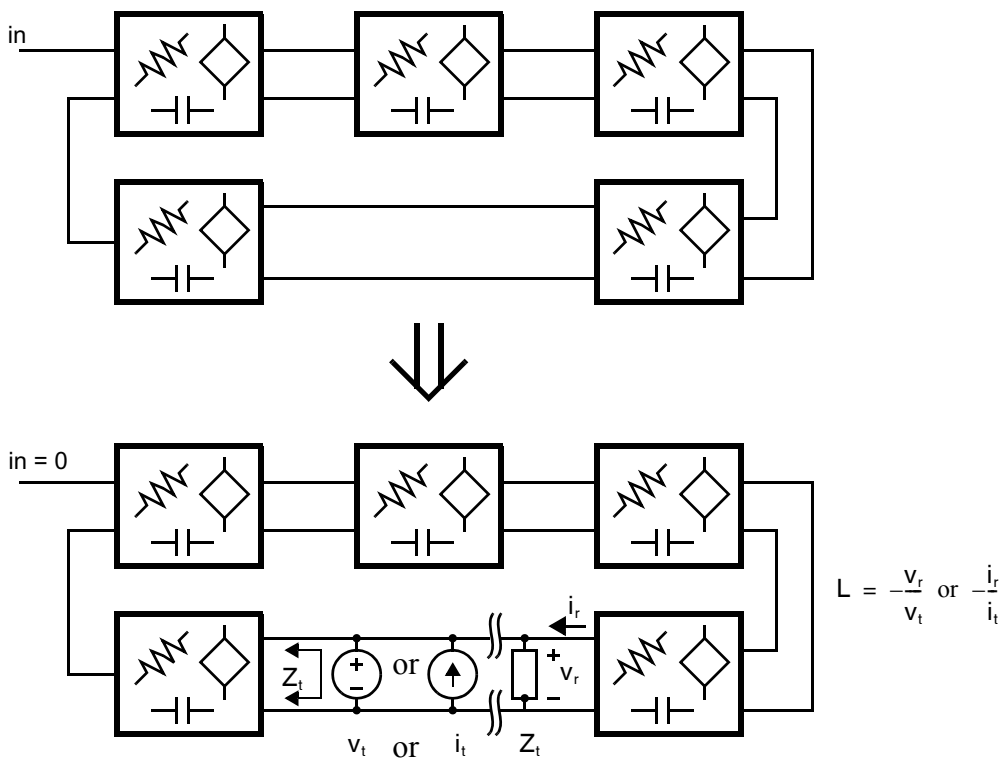


Fig. 5.11 Determining the loop gain, L, by breaking the loop.

The key step is 2. Although the loop may be broken anywhere, it is convenient to do so right at an ideal voltage or current source so that the terminating impedance plays no role in the result, and hence may be omitted. Alternatively, one may chose a point in the loop where the terminating impedance is easily found.

Key Point: *The loop gain, $L(\omega)$, may be estimated by zeroing the input, breaking the loop, and injecting a test signal. This is an accurate approximation in the typical case where high forward gain around the loop is being traded for increased bandwidth, accuracy, etc.*

The loop gain obtained in this manner is actually an approximation since it neglects the possibility of signals circulating in the opposite direction around the loop. The approximation is accurate so long as the gain around the loop in the forward direction (clockwise in Fig. 5.11) is much greater than the gain in the opposite direction (counterclockwise in Fig. 5.11). This will generally be the case in the circuits of present interest to us where feedback is used to trade high gain for other desirable properties, as described in Section 5.1.

EXAMPLE 5.8

Find the loop gain, $L(s)$, and closed-loop gain, $A_{CL}(s)$, of the inverting opamp configuration shown in Fig. 5.10 where the load impedance, Z_L, is infinite. Assume the opamp has a voltage gain A_v, and an infinite input impedance.

Solution

The small-signal equivalent representation of the circuit is shown in Fig. 5.12(*a*). To this circuit, we apply the procedure for determining the loop gain.

1. The input source is zeroed by setting v_s to ground.
2. The loop is broken at the opamp input terminals. This point is chosen because the input impedance there is infinite, so finding Z_t is trivial.
3. The test signal v_t is injected into the new circuit, Fig. 5.12(*b*). The returned voltage is determined by nodal analysis,

$$v_r = -A_v(s)\frac{Z_1}{Z_1 + Z_2 + Z_o}v_t$$

Hence, the loop gain is

$$L(s) = -\frac{v_r}{v_t} = A_v(s)\frac{Z_1}{Z_1 + Z_2 + Z_o}$$

The normalized closed-loop gain is

$$\frac{L(s)}{1 + L(s)} = \frac{A_v(s)Z_1}{A_v(s)Z_1 + Z_1 + Z_2 + Z_o}$$

The desired gain of this circuit is $1/\beta = -(Z_2/Z_1)$, so the closed loop gain is given by (5.51),

$$A_{CL}(s) = -\frac{Z_2}{Z_1} \cdot \frac{A_v(s)Z_1}{A_v(s)Z_1 + Z_1 + Z_2 + Z_o}$$

As long as $A_v \gg (Z_1 + Z_2 + Z_o)/Z_1$, the closed-loop gain will approximately equal the desired gain. The same result may be obtained by direct nodal analysis of the circuit in Fig. 5.12(*a*) without breaking the loop.

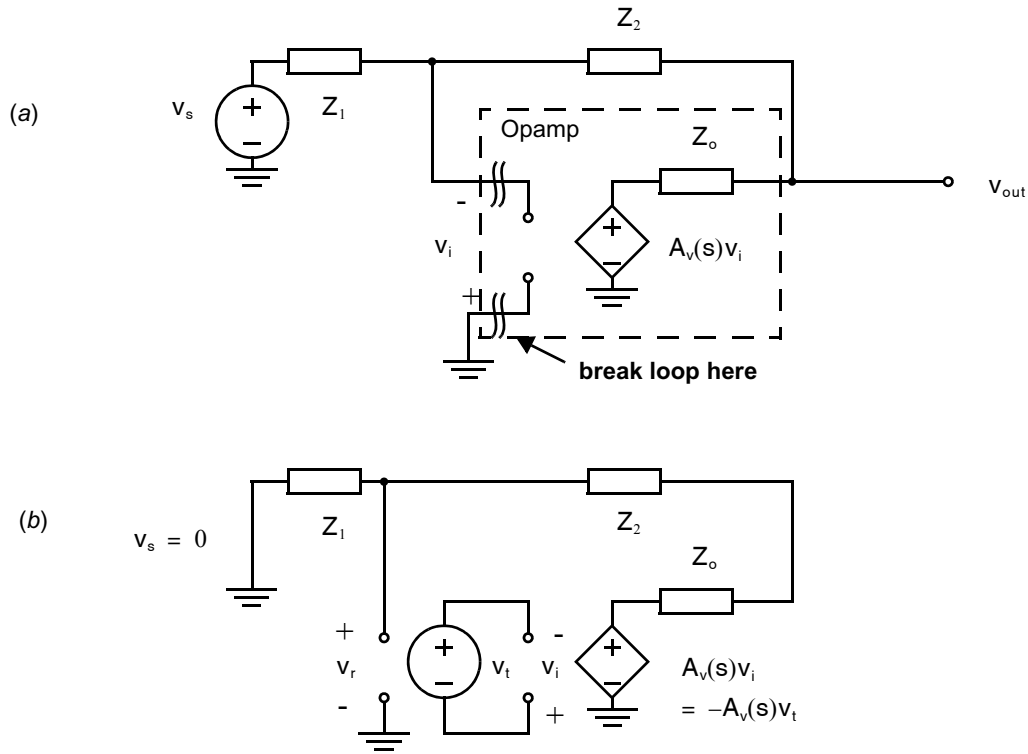


Fig. 5.12 (*a*) Small-signal model of an inverting opamp configuration. (*b*) Circuit for determining the loop gain, $L = -v_r/v_t$.

EXAMPLE 5.9

Find the loop gain of the differential pair feedback circuit shown in Fig. 5.13(*a*).

Solution

We are considering only the small-signal response of the circuit, so we first take the circuit's small-signal equivalent, shown in Fig. 5.13(*b*) where it is naturally assumed that $Q_{1,2}$ and $Q_{3,4}$ are matched device pairs. A simplified small-signal model of the differential pair with active load is adopted; for complete details the reader is referred to Chapter 3.

The circuit for determining the loop gain is shown in Fig. 5.13(*c*).

1. The input source is zeroed by setting v_{in} to ground.
2. The loop is broken at the gate of M2. This point is chosen because the terminating impedance, Z_t, is simply the input capacitance to the differential pair. In the circuit of Fig. 5.13(*c*), this terminating capacitance is placed in parallel with the (much larger) C_L, so any inaccuracy in accounting for all the parasitic components of the input capacitance will have little effect on the subsequent analysis.
3. A test source, v_t, is introduced at the break point. The circuit of Fig. 5.13(*c*) must now be analyzed to find $L = -(v_r/v_t)$. The analysis at dc follows.

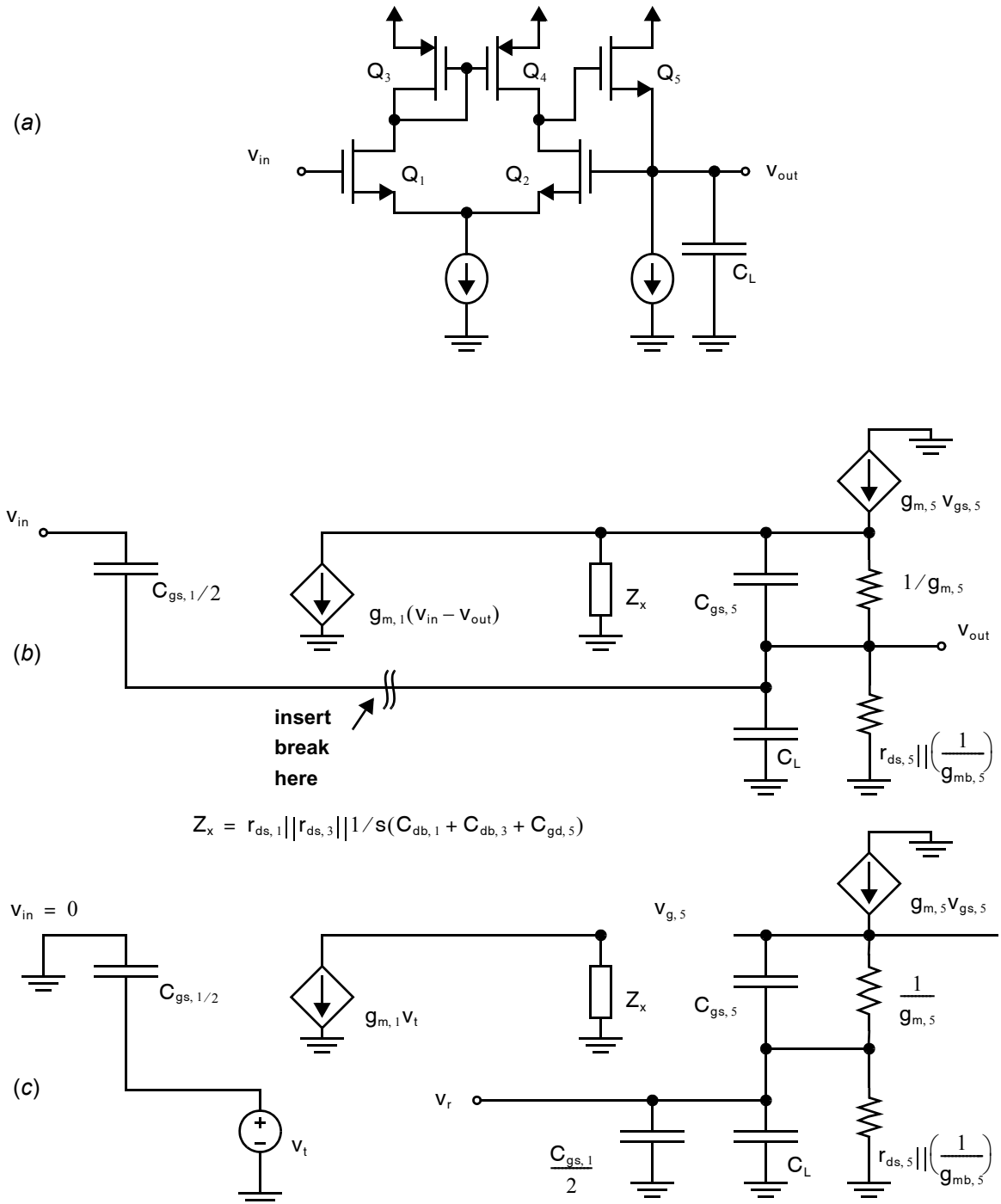


Fig. 5.13 (*a*) A unity-gain feedback circuit. (*b*) Its small-signal equivalent. (*c*) The small-signal circuit for finding the loop gain, $L = -v_r/v_t$.

Since the impedance looking into the gate of Q_5 is infinite at dc, the small-signal gate voltage is

$$v_{g,5} = -g_{m,1}v_t(r_{ds,1} \| r_{ds,3}).$$

The returned voltage is then given by a voltage division,

$$v_r = v_{g,5}\left(\frac{g_{m,5}}{g_{m,5} + 1/r_{ds,5} + g_{mb,5}}\right) = -\left[\frac{g_{m,1}(r_{ds,1} \| r_{ds,3})g_{m,5}}{g_{m,5} + 1/r_{ds,5} + g_{mb,5}}\right]v_t$$

Hence, the loop gain at dc is

$$L = \frac{g_{m,1}(r_{ds,1} \| r_{ds,3})g_{m,5}}{g_{m,5} + 1/r_{ds,5} + g_{mb,5}}$$

The high-frequency analysis is left as an exercise for the reader, where it may be useful to refer to the common-source amplifier high-frequency analysis presented in Chapter 4.

This approach is also very useful because it can be applied to computer circuit simulations to find the loop gain with great accuracy. However, the preceding steps must only be applied to the small-signal equivalent circuit. In a transistor circuit, the dc biasing connections must not be broken, otherwise some of the circuit's small-signal parameters will be completely altered. This may be achieved in simulation using a very large ideal inductance, as shown in Fig. 5.14. The large inductance is a short at dc thus preserving the dc biasing but is effectively an open-circuit at any frequencies of interest. A test small signal may then be introduced via a large coupling capacitor.

5.4.2 Noninverting Amplifier

A classic example of a feedback amplifier is the generalized noninverting opamp shown in Fig. 5.15. The operational amplifier is assumed to have a frequency-dependent voltage gain, $A_v(s)$, and complex input and output impedances given by $Z_i(s)$ and $Z_o(s)$, respectively. A complete circuit model is shown in Fig. 5.16. Note that the

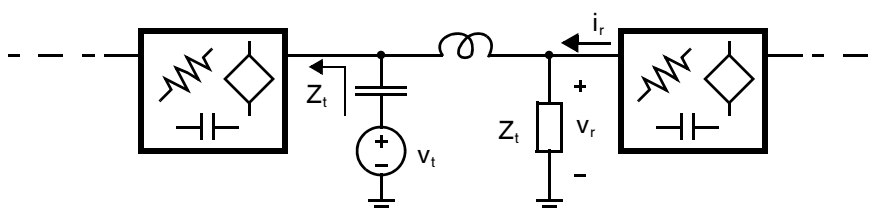


Fig. 5.14 Breaking the loop using a large ideal inductance and introducing a test signal via a large coupling capacitor to determine the small-signal loop gain in simulation without disturbing the circuit's dc biasing.

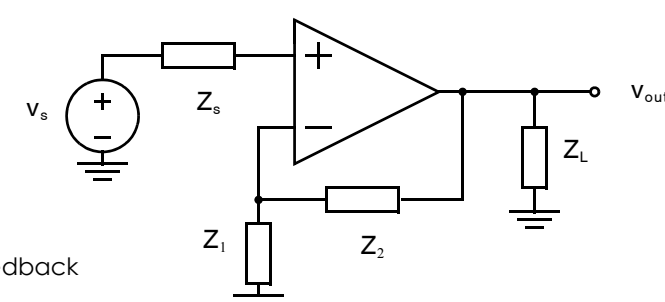


Fig. 5.15 A generalized noninverting feedback operational amplifier configuration.

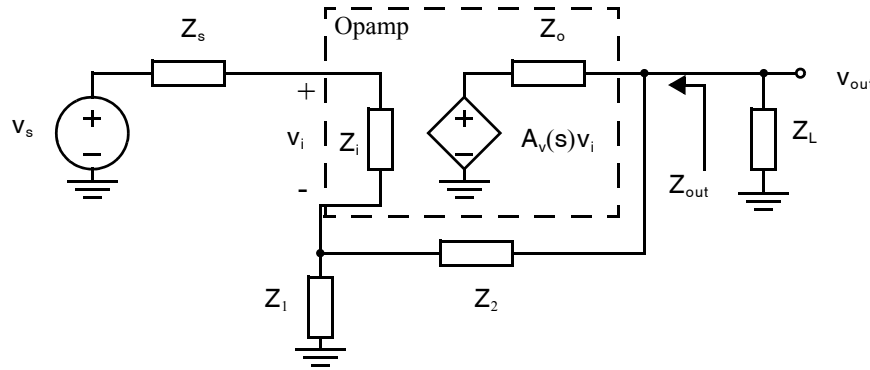


Fig. 5.16 Small-signal model of a noninverting feedback amplifier.

gain A_v is the (small-signal) open-circuit gain of the opamp and is therefore not equal to the forward-path gain A of the feedback amplifier which must include all loading.

EXAMPLE 5.10

An opamp specification indicates an input capacitance of 0.5 pF, an output resistance of 500 Ω, a dc gain of 80 dB, and a unity-gain frequency of 100 MHz. The specification also indicates that the opamp has a phase response of -105 degrees at the unity-gain frequency of $A_v(s)$. Find the opamp model components: Z_i, Z_o, and $A_v(s)$.

Solution

The input and output impedances are directly given in the specification.

$$Z_i(s) = \frac{1}{(5 \cdot 10^{-13})s}\ \Omega$$

$$Z_o(s) = 500\ \Omega$$

For the gain, we adopt a second-order model similar to (5.34).

$$A_v(s) = \frac{A_{v0}}{(1 + s/\omega_{p1})(1 + s/\omega_{eq})} \tag{5.54}$$

A dc gain of 80 dB implies $A_{v0} = 10^4$. Since the unity-gain frequency, ω_{ta}, satisfies $\omega_{p1} \ll \omega_{ta} \ll \omega_{p2}$, the dominant pole frequency is

$$\omega_{p1} = \frac{\omega_{ta}}{A_{v0}} = 2\pi\frac{10^8}{10^4} = 2\pi \cdot (10^4\ \text{Hz}). \tag{5.55}$$

To find the non-dominant pole frequency, we assume the dominant pole contributes –90° phase shift at ω_{ta}, and hence,

$$\tan^{-1}\left(\frac{\omega_{ta}}{\omega_{eq}}\right) = 15° \Rightarrow \frac{\omega_{ta}}{\omega_{eq}} = 0.270 \Rightarrow \omega_{eq} = \frac{\omega_{ta}}{0.270} = 2\pi \cdot (3.70 \cdot 10^8\ \text{Hz}) \tag{5.56}$$

It is well known that the gain of the noninverting amplifier under ideal conditions of infinite opamp gain is

$$\frac{1}{\beta} = \frac{Z_1 + Z_2}{Z_1}$$
$$\Rightarrow \beta = \frac{Z_1}{Z_1 + Z_2} \tag{5.57}$$

In order to determine the loop gain, we must analyze the circuit in Fig. 5.17. The loop is broken at the output of the controlled-source A_v, which models the opamp voltage gain. Since the termination impedance, Z_t, is inserted across an ideal voltage source, its value will have no effect on the analysis. The redrawn schematic at the bottom of Fig. 5.17 clearly shows

$$v_r = A_v(s)v_i = -A_v(s)\frac{Z_i}{Z_o}\left[\frac{Z_L||Z_o}{Z_L||Z_o + Z_2 + Z_1||(Z_s + Z_i)}\right]\left[\frac{Z_1}{Z_s + Z_i + Z_1}\right]v_t$$
$$L(s) = A_v(s)\frac{Z_i}{Z_o}\left[\frac{Z_L||Z_o}{Z_L||Z_o + Z_2 + Z_1||(Z_s + Z_i)}\right]\left[\frac{Z_1}{Z_s + Z_i + Z_1}\right] \tag{5.58}$$

A bode plot of $L(s)$ gives the phase margin. Substituting (5.57) and (5.58) into (5.51) and (5.52) will give the closed-loop frequency response and gain error. If the simplifying assumptions $Z_i \gg Z_1, Z_s$ and $Z_o \ll Z_L, Z_2$ are made (i.e., the opamp has relatively high input impedance and low output impedance) (5.58) reduces to a simple result.

$$L(s) \approx A_v(s)\frac{Z_1}{Z_1 + Z_2} = A_v(s)\beta \tag{5.59}$$

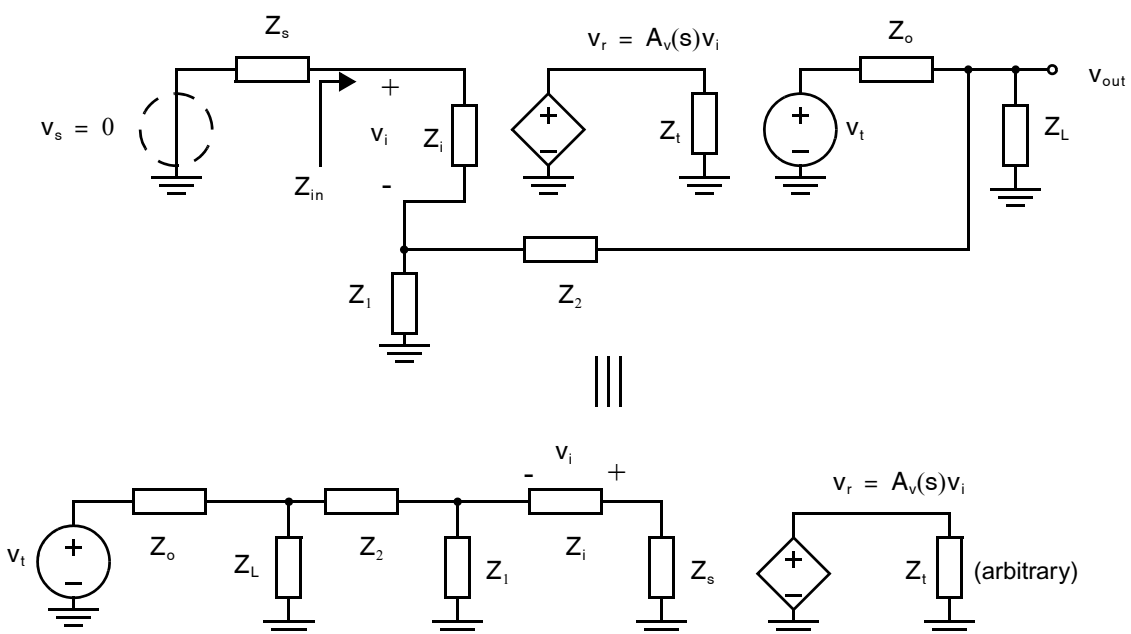


Fig. 5.17 Equivalent circuits for finding the loop gain, L(s), of a voltage feedback amplifier.

EXAMPLE 5.11

The opamp from Example 5.10 is to be used to drive a load of $Z_L = 200\Omega$ in a unity-gain configuration, as shown in Fig. 5.18. Find the dc gain error, phase margin, and bandwidth.

Solution

A unity-gain opamp configuration is a special case of the noninverting operational amplifier where $Z_1 = \infty$ and $Z_2 = 0$. Substituting these and $Z_s = 0$ into (5.58), and assuming $Z_L \ll Z_i$,

$$L(s) = A_v(s)\left(\frac{Z_L}{Z_L + Z_o}\right) \tag{5.60}$$

Since the desired gain is $1/\beta = 1$, the closed loop gain is

$$A_{CL}(s) = \frac{L(s)}{1 + L(s)} = \frac{A_v(s)Z_L}{A_v(s)Z_L + Z_L + Z_o} \tag{5.61}$$

A numerical result for the gain error is obtained using the parameter values for the opamp at dc,

$$1 - \frac{L_0}{1 + L_0} = 1 - \frac{A_{v0}Z_L}{A_{v0}Z_L + Z_L + Z_o} = 1 - \frac{10^4 200}{10^4 200 + 200 + 500} = 0.0003 \tag{5.62}$$

indicating a gain error of 0.03%.

Substituting (5.54) into equation (5.60) gives the loop gain,

$$L(s) = \frac{A_{v0}}{\left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{eq}}\right)}\left(\frac{Z_L}{Z_L + Z_o}\right) \tag{5.63}$$

A bode plot of (5.63) indicates a phase margin of 85.6°.

Finally, the bandwidth may be found by solving for the -3dB frequency of (5.61). To simplify the analysis, notice that the phase margin is close to 90° so the equivalent second pole ω_{eq} is playing only a small role at frequencies around the closed-loop bandwidth. Hence, a first-order approximation may be used for the opamp, $A_v(s) \cong \omega_{ta}/s$. Substituting this into (5.61) results in

$$A_{CL}(s) \cong \frac{1}{1 + \left(\frac{Z_L + Z_o}{Z_L}\right)\frac{s}{\omega_{ta}}} \tag{5.64}$$

from which it straightforwardly follows that the 3-dB bandwidth is approximately $(\omega_{ta}Z_L)/(Z_L + Z_o)$ which in this case equals $2\pi \cdot 2.86$ MHz.

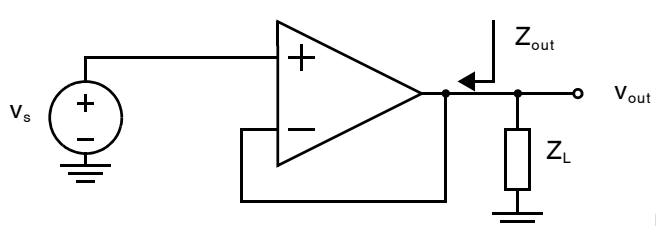


Fig. 5.18 Unity-gain opamp configuration.

In Example 5.11, an amplifier with an output impedance of 500Ω is used to drive a load of only 200Ω, which one might think would cause a reduction in the amplifier's gain. However, a result very close to the desired gain of unity is obtained. The voltage division between the opamp's output impedance and load, $Z_L/(Z_L + Z_o)$, does indeed show up as a term in (5.60) that reduces the open-loop gain, L. But, so long as the opamp has large voltage gain, A_v, the open-loop gain L remains much larger than unity and the circuit continues to work as an ideal voltage buffer with a closed-loop gain $A_{CL} \cong 1$.

EXAMPLE 5.12

What is the output impedance of the closed-loop amplifier in Example 5.11?

Solution

If the load of the closed-loop amplifier, Z_L, were equal to its output impedance, Z_{out} in Fig. 5.18, the closed-loop amplifier gain would be exactly one-half the value obtained with $Z_L = \infty$. In this case, the closed-loop gain with $Z_L = \infty$ is almost precisely unity. Hence, we can substitute $Z_L = Z_{out}$ and $A_{CL} = 1/2$ into (5.61) and solve for Z_{out}.

$$\frac{1}{2} = \frac{A_v Z_{out}}{A_v Z_{out} + Z_{out} + Z_o} \Rightarrow Z_{out} = Z_o \cdot \frac{1}{A_v - 1}$$

Substituting in dc values from Example 5.10, $Z_o = 500\Omega$ and $A_v = A_{v0} = 10^4$, gives a dc closed-loop output impedance of

$$Z_{out} = \frac{(500\Omega)}{(10^4 - 1)} \cong 0.05\Omega$$

Key Point: *Another important feature of feedback for CMOS integrated circuits is that it reduces the output-impedance of an amplifier by a factor approximately equal to the loop gain.*

In general, feedback in integrated circuits provides an effective reduction in the output impedance of an amplifier. Specifically, the output impedance of the amplifier, including the loading of the feedback network, is reduced by a factor approximately equal to the loop gain, L. This is an important feature of feedback, particularly in CMOS integrated circuits where low output impedance would otherwise require high power dissipation.[6]

Noninverting amplifiers are said to employ voltage feedback because the signal being subtracted from the source is a voltage—in this case, the voltage at the negative opamp terminal. This has the effect of increasing the amplifier's input impedance; the feedback ensures the opamp input terminals are exposed to only very small differential voltages, v_i, so clearly very little input current is required. However, this is not a particularly exciting property since MOSFET gates provide relatively high input impedance even without feedback. However, we shall see in the next section that current feedback can be used to realize very low input impedances, which can be very useful when designing a CMOS circuit to amplify current signals.

6. For example, a common-source amplifier has an output impedance that depends on its small-signal r_{ds} in active mode, which is in turn inversely proportional to its drain current. Hence, to decrease the output impedance, the drain current must increase.

5.4.3 Transimpedance (Inverting) Amplifiers

The inverting opamp configuration shown in Fig. 5.10 may actually be thought of as a transimpedance amplifier. It is redrawn in Fig. 5.19 with a Norton-equivalent source. Transimpedance amplifiers employ current feedback: the branch current through Z_f is subtracted from current coming from the source. We shall see that this results in a low input impedance, Z_i, which is useful when the input is a current. Hence, transimpedance amplifiers are useful when there is a high source impedance, Z_s. For example, they are used as amplifiers following sensors such as microphones and photodetectors.

The desired gain of the transimpedance amplifier, $1/\beta$, is the gain v_{out}/i_s obtained under conditions of very high loop gain. Consider the case where the voltage gain of the amplifier, $A_v = v_{out}/v_i$, is very large as shown in Fig. 5.20. Assuming the loop is stable v_{out} remains finite, so $v_i \cong 0$. It follows that all of the current i_s must flow through Z_f and the output voltage is $v_{out} \cong -Z_f i_s$. Hence, the desired gain for the transimpedance amplifier is

$$\frac{1}{\beta} = -Z_f \tag{5.65}$$

Notice that the circuit has a current input and voltage output, so its gain is expressed in ohms.

The loop gain is obtained, as usual, by breaking the loop and injecting a test source. Note that once the source i_s is zeroed, the transimpedance amplifier in Fig. 5.19 is topologically identical to the noninverting feedback amplifier with sources zeroed in Fig. 5.17. Hence, the loop gain is the found by making the following substitutions into (5.58):

- replace Z_1 with Z_f
- replace Z_2 with Z_s
- replace Z_s with 0

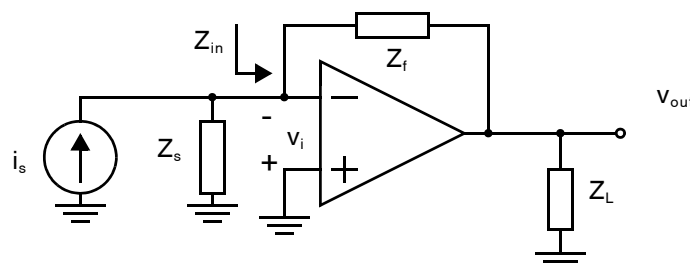


Fig. 5.19 A generalized transimpedance amplifier.

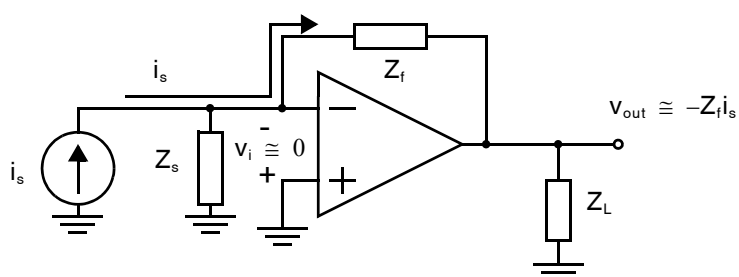


Fig. 5.20 A transimpedance amplifier with a very high voltage-gain amplifier, $A_v \to \infty$ and hence very high loop gain, L.

The results are summarized below where, as before, Z_i and Z_o are the input and output impedances of the opamp respectively.

$$L(s) = A_v(s)\frac{Z_i}{Z_o}\left[\frac{Z_L||Z_o}{Z_L||Z_o + Z_f + Z_s||Z_i}\right]\left[\frac{Z_s}{Z_i + Z_s}\right] \tag{5.66}$$

If it is assumed that $Z_i \gg Z_s$ and $Z_o \ll Z_L, Z_f$, then (5.66) simplifies considerably.

$$L(s) \approx A_v(s)\frac{Z_s}{Z_f + Z_s} \tag{5.67}$$

Notice that (5.67) is identical to (5.59) for the noninverting amplifier with $Z_s = Z_1$ and $Z_f = Z_2$. This is satisfying since the undriven circuits look nearly identical for the noninverting and transimpedance feedback amplifiers (with the exception of an additional impedance at the positive amplifier terminal in the noninverting schematics of Fig. 5.15, Fig. 5.16, and Fig. 5.17).

EXAMPLE 5.13

In the inverting feedback amplifier of Fig. 5.9, adopt a first-order model for the opamp, as in Fig. 5.21 where the input has been zeroed. Show that the system is 1st order and find the closed-loop pole.

Solution

The open-loop gain of the circuit in Fig. 5.21 may be obtained by breaking the loop at the opamp input terminals and injecting a test signal.

$$v_r = -g_m v_t\left[R_o||\left(\frac{1}{sC_1} + \frac{1}{sC_2}\right)\right]\cdot\frac{C_2}{C_1 + C_2}$$

$$\Rightarrow L(s) = -\frac{v_r}{v_t} = \frac{g_m R_o C_2}{C_1 + C_2 + sC_1C_2R_o}$$

The poles of the closed-loop amplifier are given by the solutions of $1 + L(s) = 0$.

$$1 + \frac{g_m R_o C_2}{C_1 + C_2 + sC_1C_2R_o} = 0$$

$$\Rightarrow s = -\frac{(g_m R_o C_2 + C_1 + C_2)}{C_1C_2R_o} \tag{5.68}$$

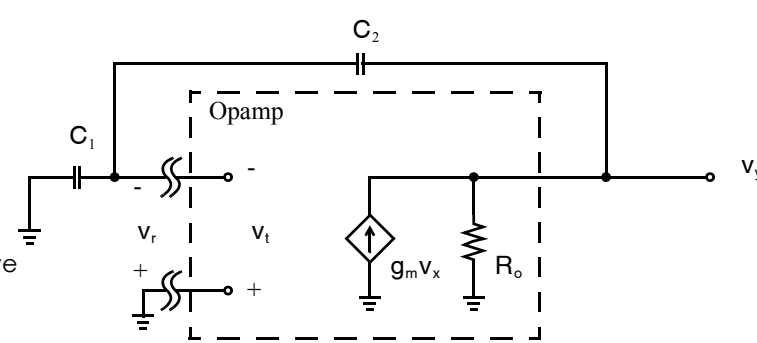


Fig. 5.21 Noninverting capacitive feedback amplifier with simple opamp model and zero input.

Hence, it is a first-order system with the stable pole on the negative real axis at $-(g_m R_o C_2 + C_1 + C_2)/(C_1 C_2 R_o)$. If one assumes $g_m R_o \gg 1$ and $C_2 \gg C_1$, we have a pole frequency of g_m/C_1.

Note the circuit poles can also be found by performing a direct nodal analysis of Fig. 5.21 with the loop closed (i.e., $v_r = v_t$) and solving the resulting system of equations for s. The node equation at the negative opamp input terminal is

$$-sC_1 v_t - sC_2(v_t + v_y) = 0 \Rightarrow v_t = -\frac{C_2}{C_1 + C_2} v_y \tag{5.69}$$

The node equation at v_y is

$$sC_2(v_y + v_t) - g_m v_t + \frac{v_y}{R_o} = 0 \tag{5.70}$$

Substituting (5.69) into (5.70) and rearranging gives the same unique pole as in (5.68).

EXAMPLE 5.14

In Fig. 5.22 a transimpedance amplifier is realized by replacing the opamp of Fig. 5.19 with a cascade of three common-source amplifiers. Find the loop gain, closed-loop gain, and input resistance, all at dc.

Solution

Assuming the transistors are in saturation, the small-signal equivalent circuit is shown in Fig. 5.23(*a*). Breaking the loop immediately after the current source, $g_m v_{gs}$, makes the analysis insensitive to the choice of the termination impedance, Z_t. Setting $i_s = 0$, introducing the test source i_t, and rearranging the schematic results in Fig. 5.23(*b*) from which the loop gain is derived.

$$L(0) = -\frac{i_r}{i_t} = g_{m,1} r_{ds,1} g_{m,2} r_{ds,2} g_{m,3} \left(\frac{r_{ds,3}}{r_{ds,3} + R_f + R_s} \right) R_s$$

The desired gain is given by (5.65). In this case

$$1/\beta = -R_f$$

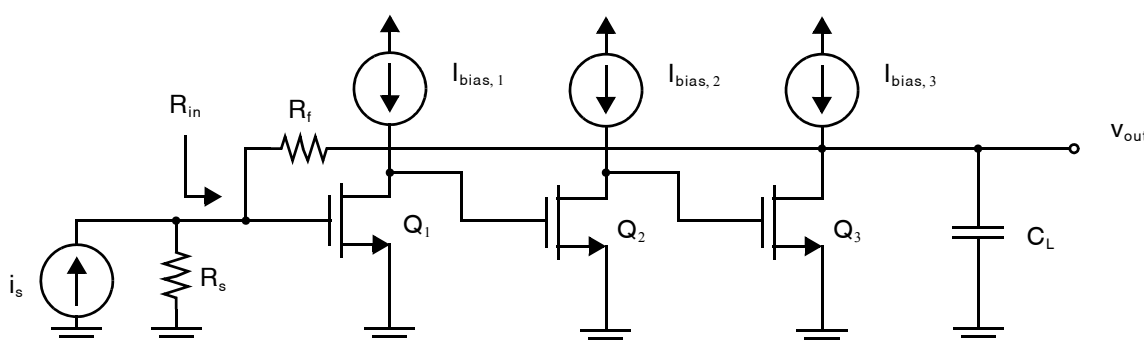


Fig. 5.22 A transimpedance amplifier based on NMOS common-source stages.

Hence, the dc closed-loop gain is

$$A_{CL}(0) = \frac{1}{\beta} \cdot \frac{L(0)}{1 + L(0)} = -R_f \cdot \left(\frac{g_{m,1} r_{ds,1} g_{m,2} r_{ds,2} g_{m,3} r_{ds,3} R_s}{g_{m,1} r_{ds,1} g_{m,2} r_{ds,2} g_{m,3} r_{ds,3} R_s + r_{ds,3} + R_f + R_s} \right) \tag{5.71}$$

which very closely approaches the desired $-R_f$ so long as each transistor's intrinsic gain $g_m r_{ds}$ is significant and R_s is not too small.

The input resistance is given by the value of R_s for which $A_{CL}(0) = -R_f/2$. Assuming $g_{m,1} r_{ds,1} g_{m,2} r_{ds,2} g_{m,3} r_{ds,3} \gg 1$ in (5.71), the result is

$$R_{in} \cong \frac{r_{ds,3} + R_f}{g_{m,1} r_{ds,1} g_{m,2} r_{ds,2} g_{m,3} r_{ds,3}} \tag{5.72}$$

For example, if $R_f = 250\Omega \ll r_{ds,3}$ and $g_m r_{ds} = 10$, then $R_{in} \cong 0.25\Omega$. Finally, note that the circuit is a third-order feedback system so there must be a dominant pole to ensure stability with phase margin, likely at the output node.

The input resistance of the transimpedance amplifier can be very small, even if the value of the feedback impedance, Z_f, is large to provide high transimpedance gain. This is because the output of the amplifier draws additional current from the input through Z_f. In general, increasing the loop gain reduces the input impedance proportionately.

Key Point: *If current feedback is used, the input-impedance of the closed-loop amplifier is reduced compared to that of the amplifier without feedback by a factor approximately equal to the loop gain.*

A low input impedance is desirable when the amplifier is driven by a current signal source. But, since any linear source may be modeled by its Thevenin/Norton equivalents, what distinguishes a voltage signal source from a current signal source? In order to observe a current signal, one wishes to have an input impedance lower than the source impedance. Hence, when the source impedance is high it is sensible to consider the input a current signal, and a low input impedance ensures little of the available source current, i_s, is lost in Z_s. However, when the source impedance is low, it is easier to consider the input a voltage signal and provide a higher input impedance.

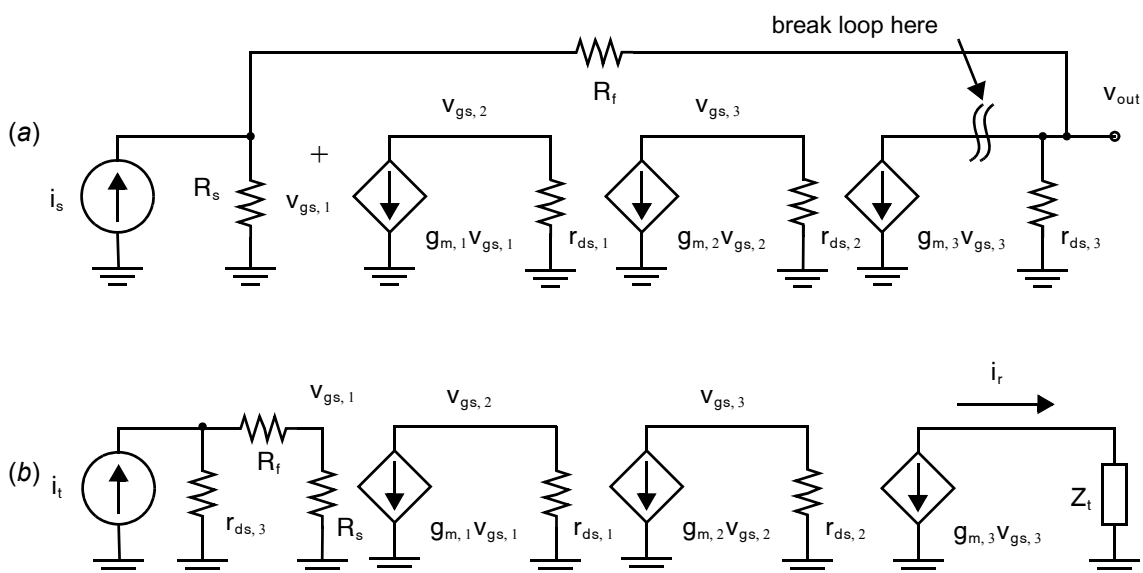


Fig. 5.23 (*a*) Small-signal model of the transimpedance amplifier of Fig. 5.22. (*b*) Loop broken and test signal introduced to determine the loop gain, $L(s)$.

5.5 SUMMARY OF KEY POINTS

- When $L>>1$, the closed-loop system has an overall gain inversely proportional to the gain of the feedback circuit and highly insensitive to the gain of the amplifier, A. [p. 206]
- A rough estimate of the bandwidth of a feedback amplifier is given by the frequency where $|L(\omega)| = 1$, which is much higher than the bandwidth of $A(\omega)$. [p. 207]
- With feedback, an amplifier is exposed to signals that are approximately L times smaller than would be the case without feedback, making it much more linear. [p. 207]
- The stability of a feedback amplifier may be checked by looking at a Bode plot of the loop gain, $L(\omega)$; so long at the phase response is greater than –180 degres at the unity gain frequency, the feedback system is stable. [p. 210]
- It is not enough that an amplifier simply be reliably stable—it must also reliably have large phase margin. To avoid undesirable behavior in the frequency and step responses of the closed-loop amplifier, it is common to require phase margins of 45 to 90 degrees. [p. 212]
- A first-order feedback amplifier is unconditionally stable with approximately 90 degrees phase margin. [p. 214]
- The closed-loop bandwidth of a feedback amplifier is ω_t, independent of the amplifier's 3-dB bandwidth, ω_{p1}. The settling time is $1/\omega_t$. [p. 216]
- All-pole second-order feedback systems are unconditionally stable with a phase margin of between 90 and 0 degrees. [p. 217]
- The process of designing or modifying a feedback amplifier to have some targeted phase margin is called compensation. [p. 218]
- Higher-order feedback amplifiers may be modeled using a second-order transfer function so long at they are not approaching instability. The second pole is chosen as the frequency at which a phase response of –135 degrees is attained. [p. 220]
- Feedback systems may be analyzed in terms of their desired gain, β, and loop gain, $L(\omega)$, from which the closed-loop amplifier's gain, bandwidth, phase margin, and frequency response may all be inferred. [p. 221]
- The loop gain, $L(\omega)$, may be estimated by zeroing the input, breaking the loop, and injecting a test signal. This is an accurate approximation in the typical case where high forward gain around the loop is being traded for increased bandwidth, accuracy, etc. [p. 223]
- Another important feature of feedback for CMOS integrated circuits is that it reduces the output-impedance of an amplifier by a factor approximately equal to the loop gain. [p. 230]
- If current feedback is used, the input-impedance of the closed-loop amplifier is reduced compared to that of the amplifier without feedback by a factor approximately equal to the loop gain. [p. 234]

5.6 REFERENCES

E. M. Cherry, "Loop Gain, Input Impedance and Output Impedance of Feedback Amplifiers," *IEEE Circuits and Systems Magazine*, Vol. 8, no. 1, pp. 55–71, 2008.

P. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits,* 5th. ed. John Wiley & Sons, New York, 2009.

A. S. Sedra and K. C. Smith. *Microelectronic Circuits,* 6th ed. Holt, Rinehart & Winston, New York, 1991.

5.7 PROBLEMS

5.7.1 Section 5.1: Ideal Model of Negative Feedback

5.1 An amplifier has a dc gain of $A = 30$. It is to be used with feedback to provide a nominal gain of 3. What feedback factor, β, should be used? What is the resulting variation in closed loop gain if A varies by $\pm 5\%$?

5.2 An amplifier, A, whose gain is 80 dB $\pm$20 dB, is enclosed in the ideal negative feedback loop of Fig. 5.1. Assuming the feedback factor, β, may be set with perfect accuracy, what is the largest closed-loop gain A_{CL} that can be realized with better than 1% accuracy?

5.7.2 Section 5.2: Dynamic Response of Feedback Amplifiers

5.3 Consider an amplifier with a transfer function

$$A(s) = \frac{10^3}{(1 + s/10^4)(1 + s/10^7)}$$

What is the approximate phase margin if this amplifier is to be used with feedback to realize a closed-loop amplifier with a gain of 10?

5.4 Spice simulations with $\beta = 1$ yield the Bode plots in Fig. P5.4 for $L(s)$. For each Bode plot:

a. Estimate the open-loop dominant pole frequency, ω_{p1}, and equivalent second pole, ω_{eq}.

b. Only the feedback network is changed resulting in a closed-loop gain of 10 V/V. Find the resulting phase margin and estimate the percentage overshoot in the closed-loop step response.

c. Only the feedback network is changed resulting in a closed-loop gain of 3 V/V. Find the 3-dB bandwidth of the resulting closed-loop amplifier?

d. If the amplifier is to be used with an adjustable feedback network to provide closed-loop gains ranging from 1 V/V to 100 V/V, is compensation required to assure a phase margin of at least 70°?

5.5 For each of the following transfer functions, A(s), sketch the Bode plots. Will they be stable or unstable when used with negative feedback for a closed-loop gain of 1 (i.e. $\beta = 1$)?

a. $A_1(s) = \dfrac{10^{11}(s + 10^7)}{(s + 10^4)(s + 10^9)}$

b. $A_2(s) = \dfrac{-10^9(s - 10^7)}{(s + 10^4)(s + 10^9)}$

5.6 An opamp with a unity gain frequency of $\omega_{ta} = 2\pi \cdot 100$ MHz is to be used within the capacitive feedback inverting configuration of Fig. 5.9. What is the maximum closed-loop gain achievable while maintaining a bandwidth of 5 MHz?

5.7.4 Section 5.3: First- and Second-Order Feedback Systems

5.7 Consider a first-order feedback system modeled by Fig. 5.1 with $A(s)$ as in (5.22) and $\omega_{p1} = 2\pi 10^4$. It is required that the closed-loop amplifier have a dc gain of 10.

a. What is the value of β?

b. Choose a value for A_0 so that the closed-loop dc gain varies by only 0.1% under $\pm 20\%$ variations in A.

c. What is the resulting 3-dB bandwidth of the closed-loop amplifier?

d. What is the resulting phase margin?

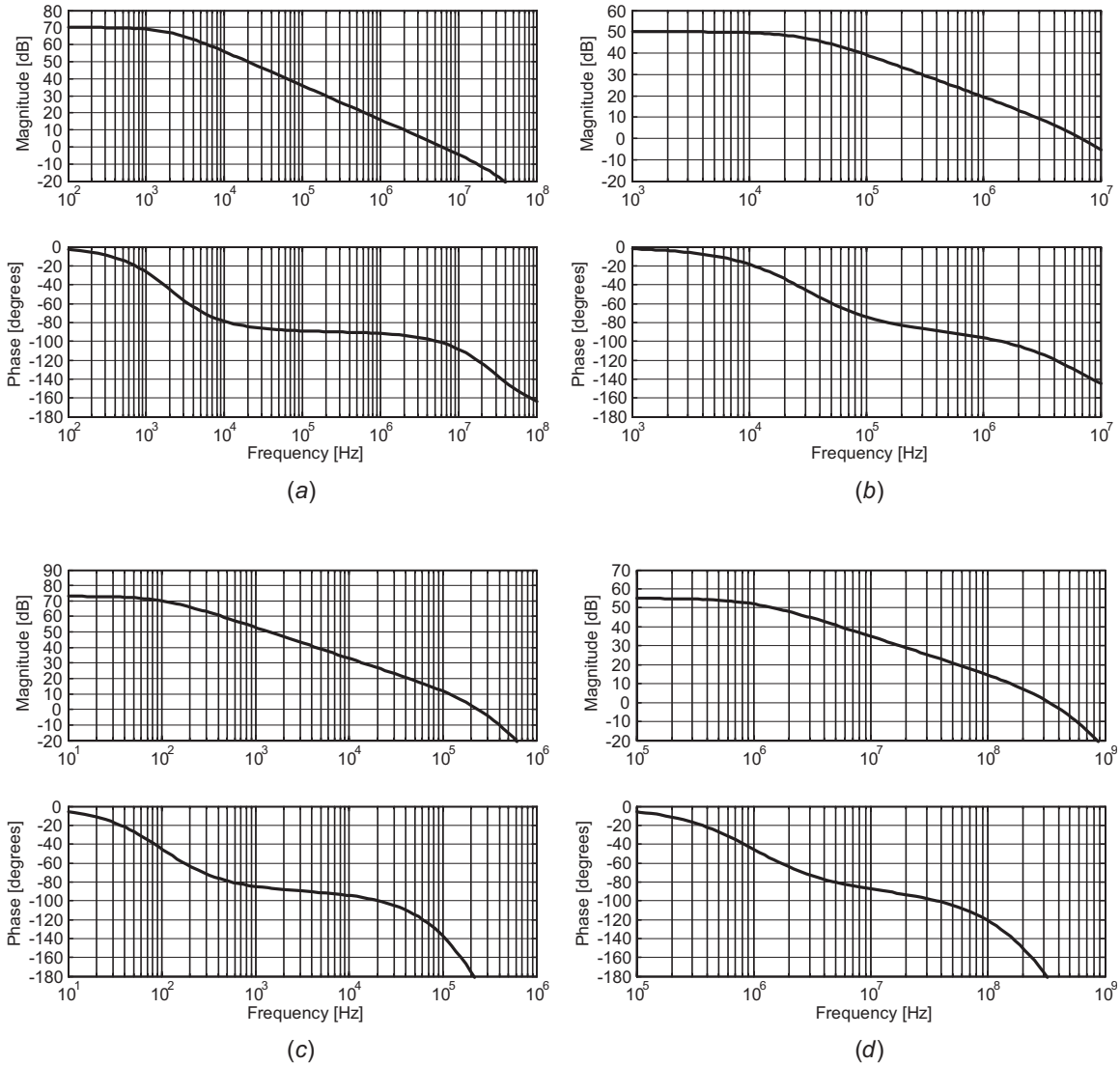


Fig. P5.4

5.8 When operated open loop, an amplifier A(s) is modeled by (5.34). Its 3 dB bandwidth is $\omega_{p1} = 2\pi 100$ kHz and its dc gain, A_0, is 80 dB. When connected with negative feedback for unity gain (i.e. $\beta = 1$), the closed loop amplifier is stable with 45° phase margin. The amplifier is now to be used with negative feedback for a closed-loop dc gain of 20 dB.

a. What is the resulting bandwidth of the closed-loop amplifier?

b. What is the resulting phase margin?

c. If ω_{p1}, ω_{p2}, and A_0 can each vary (independently) by ±20%, what are the minimum bandwidth and phase margin that can be guaranteed for the closed-loop feedback amplifier with 20 dB of dc gain?

5.9 A zero is somehow introduced into the amplifier of Problem 5.8 so that

$$A(s) = \frac{10^3(1 + s/\omega_z)}{(1 + s/10^4)(1 + s/10^7)}$$

a. What is the approximate phase margin if this amplifier is to be used with feedback to realize a closed-loop amplifier with a gain of 10 and $\omega_z = 10^7$?

b. Repeat part (*a*) with $\omega_z = -10^7$?

c. What can you conclude about the impact of a left-half plane and right-half plane zeros on the stability of feedback systems?

5.10 A feedback amplifier has a closed-loop desired gain of 2, a dc gain error of 1%, a closed-loop 3-dB bandwidth of 1 MHz, and a step response with 1% overshoot. If the open-loop amplifier $A(s)$ has a second-order response all-pole response, find the dc gain.

5.11 The amplifier shown in Fig. P5.11 is formed by cascading two first-order feedback systems. Both amplifiers have the same transfer function, $A(s)$, given by (5.22) with $A_0 = 10^4$ and $\omega_{p1} = 10^3$. The overall cascade has a dc gain of 100. Estimate the overall bandwidth? How does this compare with the overall 3-dB bandwidth that would be obtained by implementing the entire gain of 100 in a single feedback amplifier constructed from the same amplifier, $A(s)$ (but obviously a different value of β)?

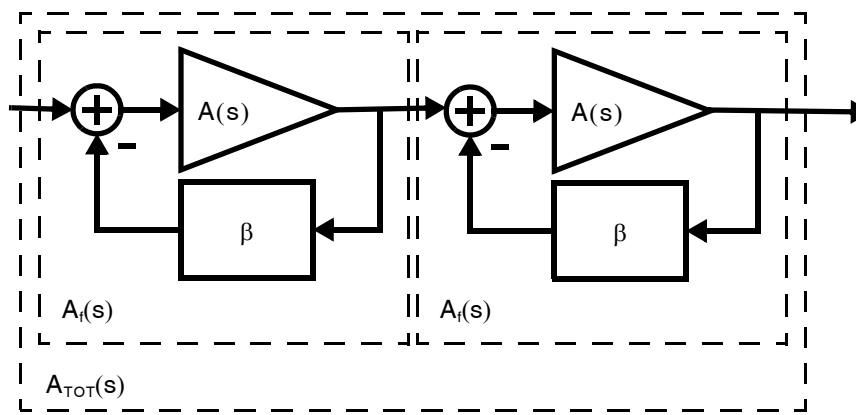


Fig. P5.11

5.12 Consider a system response, $H(s)$, with one dominant pole, and other real-axis poles, and zeros such that

$$H(s) = \frac{A_0 \times \Pi\left(1 + \dfrac{s}{\omega_{zi}}\right)}{\left(1 + \dfrac{s}{\omega_{p1}}\right)\Pi\left(1 + \dfrac{s}{\omega_{pi}}\right)}$$

Also consider an approximating function, $H_{app}(s)$, given by

$$H_{app}(s) = \frac{A_o}{\left(1 + \dfrac{s}{\omega_{p1}}\right)\left(1 + \dfrac{s}{\omega_{eq}}\right)}$$

Show that, when $H(s)$ is approximated by $H_{app}(s)$ such that the phase of the two systems are approximately equal at ω_t of $H(s)$, ω_{eq} should be equal to

$$\frac{1}{\omega_{eq}} = \sum \frac{1}{\omega_{pi}} - \sum \frac{1}{\omega_{zi}}$$

5.13 An opamp has its first pole at 3 kHz and has high-frequency poles at 130 MHz, 160 MHz, and 180 MHz. Find the frequency where the phase shift is $-135°$ and, therefore, find the equivalent time constant that models the high-frequency poles. How does this compare to the estimate given by equation (5.50)?

5.7.4 Section 5.4: Common Feedback Amplifiers

5.14 If the circuit of Fig. 5.15 is to be cast into the general model of Fig. 5.1, what voltages and/or currents play the roles of the signals u, x, v, and y? Repeat for the circuit of Fig. 5.19.

5.15 Consider the noninverting opamp in Fig. 5.15 where $Z_s = 0$, Z_1 is a resistor with value $1\text{k}\Omega$, and Z_2 is a resistor that varies over the range $0.1\text{k}\Omega$ to $10\text{k}\Omega$. The opamp may be modeled as having infinite input impedance, $Z_i = \infty$, zero output impedance, $Z_o = 0$, and a voltage gain,

$$A_v(s) = \frac{10^4}{(1 + s/10^4)(1 + s/10^8)}$$

- **a.** What is the worst-case (smallest) phase margin? What is the corresponding value of Z_2?
- **b.** What is the worst-case (lowest) bandwidth of the closed-loop amplifier? What is the corresponding value of Z_2?
- **c.** What is the worst-case (largest, slowest) time-constant of the closed-loop amplifier's step response? What is the corresponding value of Z_2?

5.16 You are given an opamp that may be modeled as in Fig. 5.16 with $Z_i = 1/s(1\text{pF})$, $Z_o = 100\Omega$, and $A_v(s) = 10$.

- **a.** The opamp is used within the unity-gain buffer in Fig. 5.18. Find the exact dc gain, approximate phase margin and closed-loop bandwidth.
- **b.** Now, the opamp is placed within the larger feedback loop depicted in Fig. P5.16(*a*). What is the dc gain and approximate phase margin?
- **c.** A capacitor, C_c, is inserted as shown in Fig. P5.16(*b*). What value is required to provide a phase margin of $90°$? What is the resulting bandwidth?
- **d.** Repeat part (*c*), but this time with the capacitor inserted as shown in Fig. P5.16(*c*).

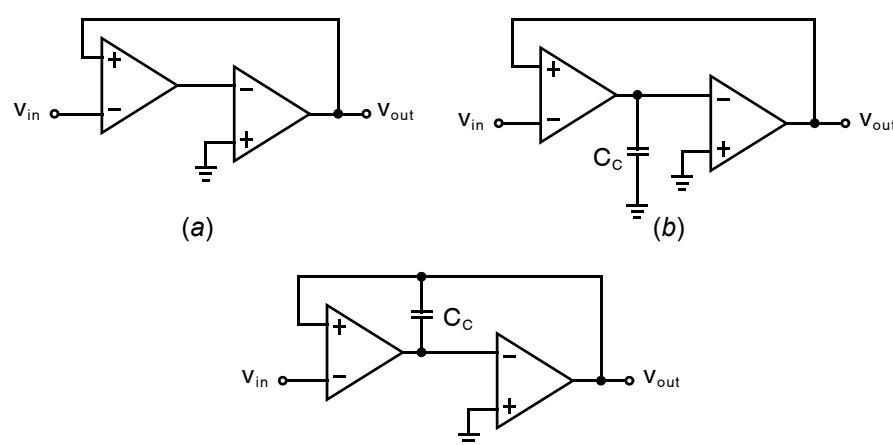


Fig. P5.16

5.17 The closed-loop system shown in Fig. P5.17(*a*) comprises three amplifiers in a loop, each modeled as in Fig. P5.17(b). What is the maximum value of A_v for which the closed-loop system is stable?

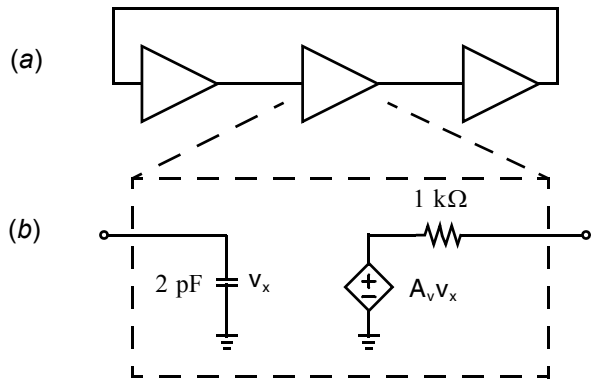


Fig. P5.17

5.18 Fig. P5.18 shows a common-gate amplifier augmented with feedback. Assume that the load resistance, $R_L < r_{ds}$. Ideally, all of the source current, i_s, passes through the transistor Q_1 and none is lost in the source resistance, R_s, resulting in a desired small-signal gain, v_{out}/i_S, of R_L.

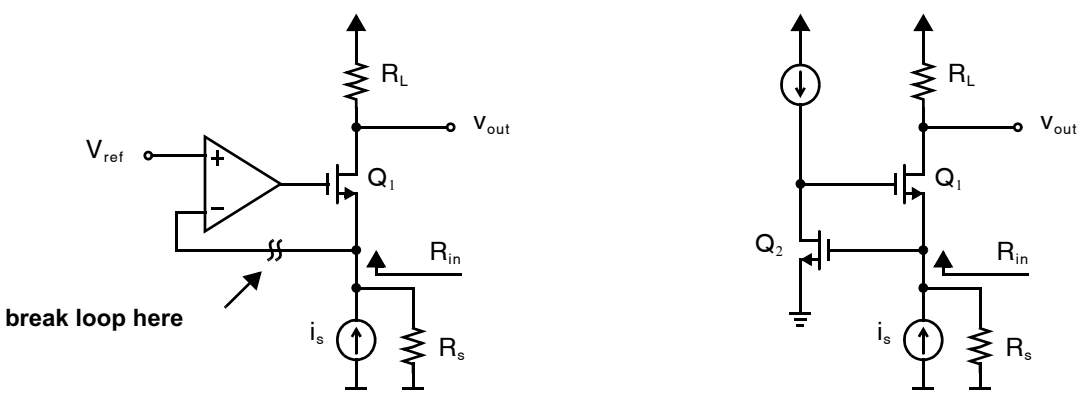


Fig. P5.18

a. What is the value of β?

b. For the circuit in Fig. P5.18(*a*), estimate the dc loop gain, L, in terms of the transistor's small-signal parameters: g_m, r_{ds}, etc. Do so by breaking the loop at the point indicated. Assume the opamp has a voltage gain A_v and infinite input impedance.

c. Combining the results in parts (*a*) and (*b*) above, write an expression for the dc gain of the closed-loop circuit, v_{out}/i_S.

d. Rewrite the expression from part (*c*) for the special case, $R_s = \infty$.

e. Solve for the source resistance, R_s, that causes the closed-loop gain in (*c*) to equal $1/2$ times the result obtained in part (d). This is a good estimate of the input resistance of the circuit, R_{in}.

f. Apply the results of parts (b) – (e) to the special case pictured in Fig. P5.18(b) where the opamp is replaced by a simple common-source amplifier.

5.19 The transimpedance amplifier in Fig. P5.19 is driven by a balanced differential input, v_{in}. Assuming all current sources are ideal and all transistors are in saturation, draw an equivalent small-signal half circuit and use it to find expressions for each of the following in terms of the transistors' small-signal parameters g_m, r_{ds}, etc.

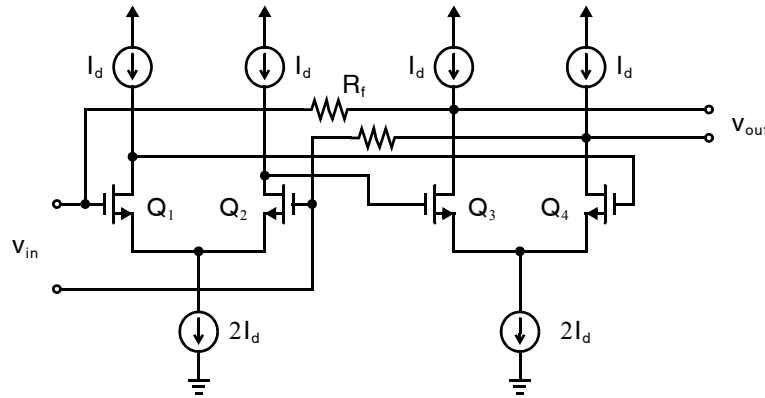


Fig. P5.19

a. Feedback factor, β.

b. Loop gain, $L(s)$.

c. Closed-loop response, $A_{CL}(s)$.

d. Input resistance at dc, R_{in}.

e. Output resistance at dc, R_{out}.

5.20 The circuit in Fig. 5.19 is to be used to provide a transimpedance gain of 10 $k\Omega$ to the current signal coming from a photodetector. The photodetector may be modeled as a current source with an impedance of $Z_s = 1\ k\Omega || 1/(s \cdot 1.5\ pF)$. Assuming the core amplifier has infinite input impedance, zero output impedance, and a first-order response, find its dc gain and unity-gain frequency ω_{ta} to ensure an overall closed-loop bandwidth of 3 GHz.

CHAPTER

6 Basic Opamp Design and Compensation

This chapter describes the fundamental principles of basic opamp design. To illustrate many of these principles, the design of a traditional opamp—namely, the two-stage CMOS opamp—is used first. This example illustrates compensation techniques needed to ensure stability in closed-loop amplifiers as well as a number of other important design techniques, such as ensuring zero systematic input-offset-voltage and process-insensitive lead compensation.

Although the two-stage CMOS opamp is a classic circuit used in many integrated circuits, other architectures have gained popularity. A few of the most important applications of opamps in analog integrated circuits are illustrated in Fig. 6.1. Note that in many cases, the loads being driven by the opamp are purely capacitive. This fact may be used to advantage by designing opamps to have high output impedances, providing large voltage gain in a single stage with relatively low power consumption. Opamps exemplifying this approach are described later in the chapter. However, one should not overlook the potential of the two-stage opamp, which is well-suited to low-voltage applications since it does not require a cascode output stage. Fully-differential opamps are also described.

6.1 TWO-STAGE CMOS OPAMP

The two-stage circuit architecture has historically been the most popular approach to opamp design. When properly designed, the two-stage opamp has a performance very close to designs that employ cascode stages and is somewhat more suitable when resistive loads need to be driven.[1] It can provide a high gain and high output swing, making it an important circuit for advanced CMOS technologies where transistor intrinsic gain and supply voltages may be limited. Furthermore, it is an excellent example to illustrate many important design concepts that are also directly applicable to other designs.

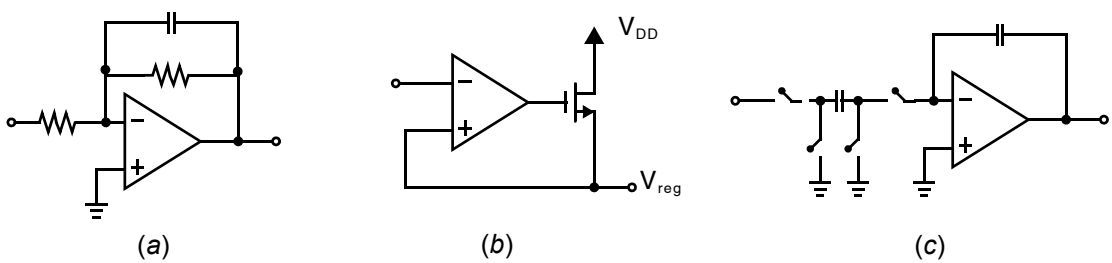


Fig. 6.1 Typical applications of opamps in analog integrated circuits: (*a*) amplification and filtering; (*b*) biasing and regulation; (*c*) switched-capacitor circuits.

1. In a CMOS integrated circuit, opamp loads are often, but not always, purely capacitive.

A block diagram of a typical two-stage CMOS opamp is shown in Fig. 6.2. "Two-stage" refers to the number of gain stages in the opamp. Fig. 6.2 actually shows three stages—two gain stages and a unity-gain output stage. The output buffer is normally present only when resistive loads need to be driven. If the load is purely capacitive, then it is seldom included. The first gain stage is a differential-input single-ended output stage, often very similar to that shown in Fig. 3.19. The second gain stage is normally a common-source gain stage that has an active load, often very similar to that shown previously in Fig. 3.4. Capacitor C_C is included to ensure stability when the opamp is used with feedback. Because C_C is between the input and the output of the high-gain second stage, it is often called a Miller capacitance since its effective capacitive load on the first stage is larger than its physical value. The third stage, when present, is typically a source-follower, similar to that shown in Fig. 3.6.

Key Point: *The classic two-stage opamp comprises a differential input gain stage, a common-source second gain stage, and optionally a unity-gain output buffer stage.*

An example of a practical CMOS version of the two-stage opamp is shown in Fig. 6.3. In this example, it is assumed that a capacitive load is being driven, so the source-follower output buffer is omitted. This example is used to illustrate many of the important design principles when realizing the two-stage amplifier.

It should be noted that the first stage has a p-channel differential input pair with an n-channel current-mirror active load. This is a complementary differential gain stage to that shown previously in Fig. 3.19. The trade-offs between having p-channel input transistors versus this stage and the alternative stage of Fig. 3.19 will be discussed later in this chapter. Also, the numbers next to the transistors represent reasonable transistor sizes for a 0.18-μm process. Reasonable sizes for the lengths of the transistor might be somewhere between 1.5 and 2 times the minimum transistor length of a particular technology, whereas digital logic typically makes use of the minimum transistor length.

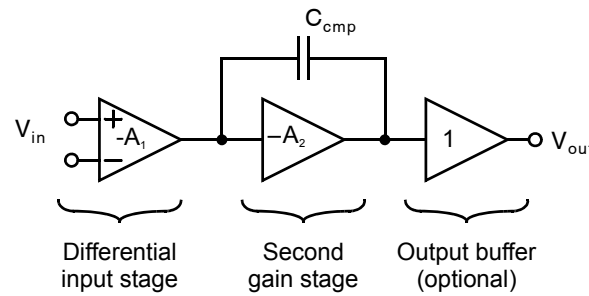


Fig. 6.2 A block diagram of a two-stage opamp.

6.1.1 Opamp Gain

First we discuss the overall gain of the opamp. For low-frequency applications, this gain is one of the most critical parameters of an opamp.

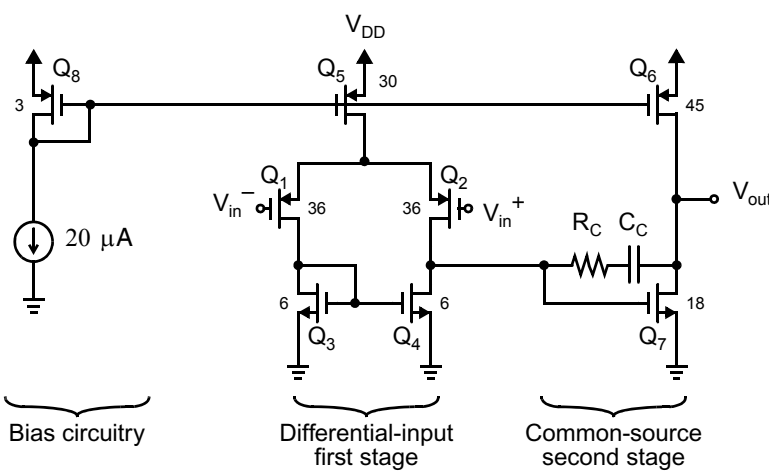


Fig. 6.3 A CMOS two-stage amplifier. All transistor lengths are 0.3 μm and widths are shown in μm next to each transistor.

The gain of the first stage has already been derived, resulting in (3.79), and is repeated here for convenience:

$$A_{v1} = -g_{m1}(r_{ds2} \| dr_{ds4}) \tag{6.1}$$

Recall from Chapter 1 that g_{m1} is given by

$$g_{m1} = \sqrt{2\mu_p C_{ox}\left(\frac{W}{L}\right)_1 I_{D1}} = \sqrt{2\mu_p C_{ox}\left(\frac{W}{L}\right)_1 \frac{I_{bias}}{2}} \tag{6.2}$$

The second gain stage is simply a common-source gain stage with a p-channel active load, Q_6. Its gain is given by

$$A_{v2} = -g_{m7}(r_{ds6} \| r_{ds7}) \tag{6.3}$$

Key Point: *The dc gain of the two-stage opamp is approximated by the product of two stages, each having a gain of approximately* $g_m r_{ds}/2$.

The third stage, if included, is a common-drain buffer stage similar to that shown in Fig. 3.6. This stage is often called a *source follower*, because the source voltage follows the gate voltage, except for a level shift. When possible, it is desirable to tie the substrate of the source-follower device to its source in order to eliminate gain degradations due to the body effect. This connection also results in a smaller dc voltage drop from the gate to the source of the source-follower device, which is a major limitation on the maximum positive output voltage.

EXAMPLE 6.1

SPICE! *Simulate this opamp with the provided netlist.*

Find the gain of the opamp shown in Fig. 6.3. Assume the power supply is $V_{DD} = 1.8$ V and a purely capacitive load. Assume the process parameters for the 0.18-μm process in Table 1.5.

Solution

First the bias currents are calculated. Since $I_{D8} = 20\ \mu A$, we have

$$I_{D1} = I_{D2} = I_{D3} = I_{D4} = I_{D5}/2 = (W_5/2W_8)I_{D8} = 100\ \mu A$$

and

$$I_{D6} = I_{D7} = (W_6/W_5)I_{D5} = 300\ \mu A$$

We can now calculate the transconductances of Q_1, Q_2, Q_7, and Q_8. Using (6.2), we have $g_{m1} = g_{m2} = 1.30$ mA/V, and $g_{m7} = 3.12$ mA/V.

Next, we estimate the output impedances of the transistors using (1.86). It should be mentioned here that (1.86) is only of very moderate accuracy, perhaps 50 percent at best. These values for the transistor output impedances can be corrected later once a SPICE analysis is run. With the use of (1.86), we find that all of the transistors in the first stage have output impedances of approximately

$$r_{ds1} = r_{ds2} = r_{ds3} = r_{ds4} = 37.5\ k\Omega$$

The output impedances of transistors in the second stage are

$$r_{ds6} = r_{ds7} = 12.5\ k\Omega$$

Finally, using (6.1), we calculate that

$$A_{v1} = -g_{m1}(r_{ds2} \| r_{ds4}) = -24.4\ V/V$$

Using (6.3),

$$A_{v2} = -g_{m7}(r_{ds6} \| r_{ds7}) = -19.5\ V/V$$

Thus, the total gain is equal to $A_{v1}A_{v2} = -1043$ V/V or 60.4 dB. Once again, it should be mentioned here that this result is a rough approximation and should be verified using SPICE. The benefit of performing the hand calculations is to see how the gain is affected by different design parameters.

6.1.2 Frequency Response

First-Order Model

We now wish to investigate the frequency response of the two-stage opamp at frequencies where the compensation capacitor, C_C, has caused the magnitude of the gain to begin to decrease, but still at frequencies well below the unity-gain frequency of the opamp. This corresponds to midband frequencies for many applications. This allows us to make a couple of simplifying assumptions. First, we will ignore all capacitors except the compensation capacitor, C_C, which normally dominates at all frequencies except around the unity-gain frequency of the opamp.[2] Second, we also assume that resistor R_C is not present. This resistor is included to achieve lead compensation and it has an effect only around the unity-gain frequency of the opamp, as we will see when we discuss compensation in Section 6.2. The simplified circuit used for analysis is shown in Fig. 6.4. It is worth mentioning that this simplified circuit is often used during system-level simulations when speed is more important than accuracy, although aspects such as slew-rate limiting (see next subsection) should also be included in the simulation.

The second stage introduces primarily a capacitive load on the first stage due to the compensation capacitor, C_C. Using Miller's theorem (Section 4.2.3), one can show that the equivalent load capacitance, C_{eq}, at node v_1 is given by

$$C_{eq} = C_C(1 + A_{v2}) \approx C_C A_{v2} \tag{6.4}$$

The gain in the first stage can now be found using the small-signal model of Figure 4.37, resulting in

$$A_{v1} = \frac{v_1}{v_{in}} = -g_{m1}Z_{out1} \tag{6.5}$$

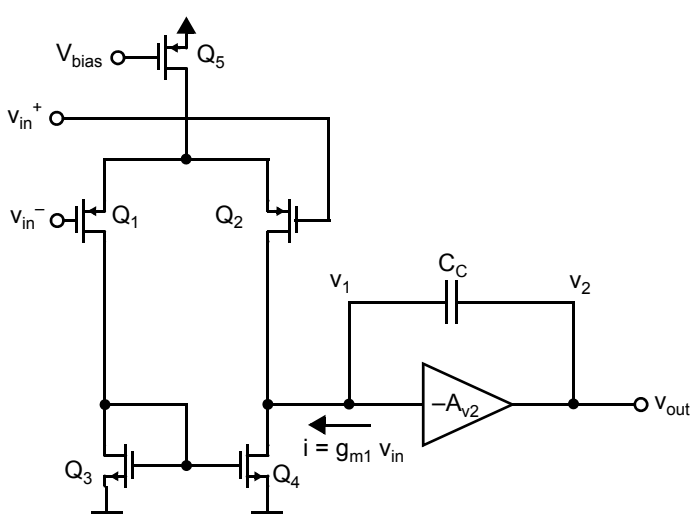


Fig. 6.4 A simplified model for the opamp used to find the midband frequency response.

2. Recall that the unity-gain frequency of an opamp is that frequency where the magnitude of the open-loop opamp gain has decreased to 1.

where

$$Z_{out1} = r_{ds2} \parallel r_{ds4} \parallel \frac{1}{sC_{eq}} \tag{6.6}$$

For midband frequencies, the impedance of C_{eq} dominates, and we can write

$$Z_{out1} \cong \frac{1}{sC_{eq}} \cong \frac{1}{sC_C A_{v2}} \tag{6.7}$$

For the overall gain, we have

$$A_v(s) \equiv \frac{v_{out}}{v_{in}} = A_{v2}A_{v1} \cong A_{v2}\frac{g_{m1}}{sC_C A_{v2}} = \frac{g_{m1}}{sC_C} \tag{6.8}$$

using (6.5) and (6.7). This simple equation can be used to find the approximate[3] unity-gain frequency. Specifically, to find the unity-gain frequency, ω_{ta}, we set $|A_v(j\omega_{ta})| = 1$, and solve for ω_{ta}. Performing such a procedure with (6.8), we obtain the following relationship:

$$\omega_{ta} = \frac{g_{m1}}{C_C} \tag{6.9}$$

Note here that the unity-gain frequency is directly proportional to g_{m1} and inversely proportional to C_C. Substituting (1.78) into (6.9) provides another expression for the unity-gain frequency.

$$\omega_{ta} = \frac{2I_{D1}}{V_{eff1}C_C} = \frac{I_{D5}}{V_{eff1}C_C} \tag{6.10}$$

Key Point: *The two-stage opamp's frequency response is dominated by a pole at the input to the second stage arising from the introduced Miller capacitor,* C_C. *A dominant-pole approximation yields a unity-gain frequency of* g_{m1}/C_C.

Equation (6.10) indicates that, for a fixed unity-gain frequency, the bias current and hence power consumption is minimized with small values of V_{eff1}. Ordinarily, a downside of selecting a small value for V_{eff1} would be increased distortion, but this is mitigated in the case of opamps used with feedback since only very small differential signals appear at the differential pair inputs. Equation (6.10) presumes a square-law for Q_1, Q_2; the current is minimized when these devices are operated in subthreshold where, from (1.122), we have $g_{m1(sub-th)} = qI_{D1}/nkT$ which, when substituted into (6.9) yields $\omega_{ta} = qI_{D1}/nkTC_C$.

EXAMPLE 6.2

Using the same parameters as in Example 6.1, and assuming $C_C = 1$ pF, what is the unity-gain frequency in Hz?

Solution

Using $g_{m1} = 1.30$ mA/V and (6.9), we find that

$$\omega_{ta} = \frac{1.3 \times 10^{-3}\ \text{A/V}}{1 \times 10^{-12}\ \text{F}} = 1.3 \cdot 10^{9}\ \text{rad/sec}$$

Thus, we find that $f_{ta} = \omega_{ta}/(2\pi) = 207$ MHz.

3. It is approximate because the higher-frequency poles are being ignored. Both of these effects will cause the unity-gain frequency to be slightly smaller than that calculated.

Second-Order Model

A simplified small-signal model for the two-stage opamp is shown in Fig. 6.5, where the output buffer has again been ignored. Also, it is assumed that any parasitic poles in the first stage are at frequencies much higher than the opamp unity-gain frequency ω_{ta} and that the first stage can therefore be modelled by a simple voltage-controlled current source. In the small-signal model, we have

$$R_1 = r_{ds4} \parallel r_{ds2} \tag{6.11}$$

$$C_1 = C_{db2} + C_{db4} + C_{gs7} \tag{6.12}$$

$$R_2 = r_{ds6} \parallel r_{ds7} \tag{6.13}$$

$$C_2 = C_{db7} + C_{db6} + C_{L2} \tag{6.14}$$

Note that if no output buffer is present, then C_{L2} is the output load capacitance. If an output buffer is present, then C_{L2} is the load capacitance introduced by the output buffer.

To show the need for R_C, we first assume $R_C = 0$ and perform nodal analysis at the nodes designated by v_1 and v_{out}. The following transfer function is obtained:

$$A_v(s) = \frac{v_{out}}{v_{in}} = \frac{g_{m1}g_{m7}R_1R_2\left(1 - \frac{sC_C}{g_{m7}}\right)}{1 + sa + s^2b} \tag{6.15}$$

where

$$a = (C_2 + C_C)R_2 + (C_1 + C_C)R_1 + g_{m7}R_1R_2C_C \tag{6.16}$$

and

$$b = R_1R_2(C_1C_2 + C_1C_C + C_2C_C) \tag{6.17}$$

It is possible to find approximate equations for the two poles based on the assumption that the poles are real and widely separated.[4] This assumption allows us to express the denominator, $D(s)$, as[5]

$$D(s) = \left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{p2}}\right) \cong 1 + \frac{s}{\omega_{p1}} + \frac{s^2}{\omega_{p1}\omega_{p2}} \tag{6.18}$$

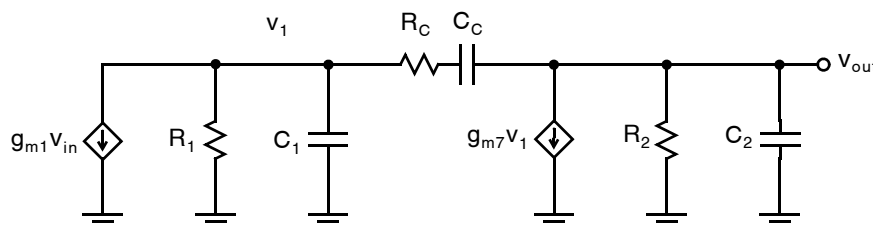


Fig. 6.5 A small-signal model of the two-stage opamp used for compensation analysis.

4. If this assumption is not valid, then it is extremely difficult to properly compensate the opamp for unity-gain stability.
5. Note that the notation $[1 + (s/\omega_x)]$ implies that the root is in the left half plane if $\omega_x > 0$, whereas the root is in the right half plane if $\omega_x < 0$.

Equating the coefficients of (6.15) equal to the coefficients of (6.18) and solving for ω_{p1} and ω_{p2} results in the following relationships. The dominant pole, ω_{p1}, is given by

$$\begin{aligned}\omega_{p1} &\cong \frac{1}{R_1[C_1 + C_C(1 + g_{m7}R_2)] + R_2(C_2 + C_C)} \\ &\cong \frac{1}{R_1C_C(1 + g_{m7}R_2)} \\ &\cong \frac{1}{g_{m7}R_1R_2C_C}\end{aligned} \tag{6.19}$$

whereas the nondominant pole, ω_{p2}, is given by

$$\begin{aligned}\omega_{p2} &\cong \frac{g_{m7}C_C}{C_1C_2 + C_2C_C + C_1C_C} \\ &\cong \frac{g_{m7}}{C_1 + C_2}\end{aligned} \tag{6.20}$$

Note that, from (6.15), another zero, ω_z, is located in the right half plane and is given by

$$\omega_z = \frac{-g_{m7}}{C_C} \tag{6.21}$$

This zero should come as no surprise since there are two signal paths having opposite signs in the small-signal circuit of Fig. 6.5—one through the voltage-controlled current source, $g_{m7}v_1$, and the other through the compensation capacitor, C_C.

Key Point: *The two-stage opamp's second pole arises at the output node and may be increased by increasing the second-stage transconductance,* g_{m7}.

From the preceding relationships for ω_{p1} and ω_{p2}, we can see that, as g_{m7} increases, the separation between the first and second poles increases. This separation tends to make the circuit more stable; hence, the use of a Miller capacitance for compensation is often called *pole-splitting compensation.* Also note that increasing C_C moves the dominant pole, ω_{p1}, to a lower frequency without affecting the second pole, ω_{p2}. This effect also makes the opamp more stable.

However, a problem arises due to the right-half-plane zero, ω_z. Because the zero is in the right half plane, it introduces negative phase shift, or phase lag, in the transfer function of the opamp. This makes stability more difficult. Making C_C larger does not help matters because this decreases the frequencies of both the first pole and the zero without making them more widely separated. Indeed, because of this right-half-plane zero, it is often impossible to choose C_C such that the step response has no overshoot (assuming $R_C = 0$). Fortunately, all is not lost, since introducing R_C allows adequate compensation, as we discuss in the next subsection.

EXAMPLE 6.3

Using the same parameters as in Example 6.1, and assuming $C_C = 1$ pF and a load capacitance $C_{L2} = 5$ pF, estimate the second pole and zero frequencies. Suggest a design change that would increase both frequencies.

Solution

Assuming C_{db6} and C_{db7} are both much smaller than C_{L2}, we take $C_2 \cong C_{L2} = 5$ pF. Substituting this and $g_{m7} = 3.12$ mA/V into (6.20),

$$\omega_{p2} \cong \frac{3.12 \text{ mA/V}}{5 \text{ pF} + 1 \text{ pF}} = 520 \cdot 10^6 \text{ rad/sec} = 2\pi \cdot 82.8 \text{ MHz}$$

The zero is located at

$$|\omega_z| \cong \frac{3.12 \text{ mA/V}}{5 \text{ pF}} = 624 \cdot 10^6 \text{ rad/sec} = 2\pi \cdot 99.3 \text{ MHz}$$

Both ω_{p2} and $|\omega_z|$ could be increased by widening both Q_6 and Q_7, which would in turn increase g_{m7}. However, doing so would also increase the current consumption of second stage.

Normally one has little control over ω_{ta}, assuming a given maximum power dissipation is allowed. It is usually constrained by stability requirements to be less than approximately one-half of the second-pole frequency, ω_{p2}, as we have seen in Section 5.3.2.[6] The second pole frequency approximated in (6.20) is, in turn, dependent upon the load capacitance via C_2, and the power dissipation of the second stage via g_{m7}.

6.1.3 Slew Rate

> ***Key Point:*** *The slew rate is the maximum rate at which the output of an opamp changes when a large differential input signal is present.*

The maximum rate at which the output of an opamp can change is limited by the finite bias currents within it. When the input to an opamp circuit changes too quickly, the opamp is unable to maintain a virtual ground between its inputs so the opamp temporarily sees a large differential input. The opamp's output voltage then changes at its maximum rate, called the *slew rate,* SR. Under slew-rate limiting, the opamp's response is nonlinear. The effect is illustrated in Fig. 6.6 where a large step input causes the output to increase at its slew rate until such time as it is able to resume linear operation without exceeding the opamp slew rate. Such transient behavior is common in switched capacitor circuits, where the slew rate is a major factor determining the circuit's settling time.

EXAMPLE 6.4

Consider a closed-loop feedback amplifier with a first-order linear settling time constant of $\tau = 0.2$ μs and a slew rate of 1 V/μs. What is the time required for the output to settle when generating to a 10-mV step output with 0.1 mV accuracy? What about a 1-V step output?

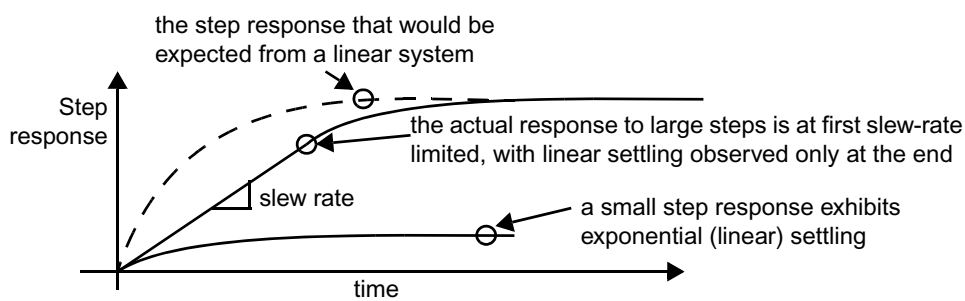

the step response that would be
expected from a linear system
Step
response
the actual response to large steps is at first slew-rate
limited, with linear settling observed only at the end
slew rate
a small step response exhibits
exponential (linear) settling
time

Fig. 6.6 An illustration of the effect of slew-rate limiting in opamps.

---

6. Assuming a unity-gain feedback configuration ($\beta = 1$ and $\omega_t = \omega_{ta}$) this will provide over $60°$ phase margin.

Solution

The amplifier will slew-rate limit whenever the slope demanded by linear settling exceeds the maximum imposed by slew-rate limiting. For a step height V_{step}, linear settling is exponential.

$$v_o = V_{step}(1 - e^{-t/\tau}) \tag{6.22}$$

The highest rate of change is observed right at the time of the step.

$$\left.\frac{dv_o}{dt}\right|_{max} = \left.\frac{dv_o}{dt}\right|_{t=0} = \frac{V_{step}}{\tau} \tag{6.23}$$

So long as this maximum slope is less than the slew rate, slew-rate limiting is avoided. Hence, the maximum step size that can be tolerated without slew-rate limiting is

$$V_{step,\,max} < SR \cdot \tau \tag{6.24}$$

In this case, linear settling is observed as long as the output step amplitude is less than 0.2 V. For a 10-mV step, the slope predicted by (6.23) is only 0.05 V/s, far below the specified maximum slew rate. Hence, 99% settling is observed after $-\ln(0.1/10)\tau = 4.6\tau = 0.92\ \mu s$.

With a 1-V step, the amplifier output will initially increase at the slew rate until it is close enough to its final value that linear settling may begin. This happens when the output is within $V_{step,\,max}$ of its final value. Hence, the amplifier will be slew-rate limited for $(1\text{ V} - 0.2\text{ V})/1\text{ V}/\mu s = 0.8\ \mu s$. From this time on, linear settling is observed and $-\ln(0.1/200) = 7.6$ additional time constants are required for settling with 0.1 mV accuracy, as shown in Fig. 6.7. Hence, the total settling time is $0.8\ \mu s + 7.6\tau = 2.32\ \mu s$. If the amplifier slew rate were made large enough that slew-rate limiting could be avoided, the settling time would be only $-\ln(0.1/1000)\tau = 1.84\ \mu s$.

When the opamp of Fig. 6.3 is limited by its slew rate because a large input signal is present, all of the bias current of Q_5 goes into either Q_1 or Q_2, depending on whether v_{in} is negative or positive. When v_{in} is a large positive voltage, the bias current, I_{D5}, goes entirely through Q_1 and also goes into the current-mirror pair, Q_3, Q_4. Thus, the current coming out of the compensation capacitor, C_C, (i.e., I_{D4}) is simply equal to I_{D5} since Q_2 is off. When v_{in} is a large negative voltage, the current-mirror pair Q_3 and Q_4 is shut off because Q_1 is off, and now the bias current, I_{D5}, goes directly into C_C. In either case, the maximum current entering or leaving C_C is simply the total bias current, I_{D5}.[7]

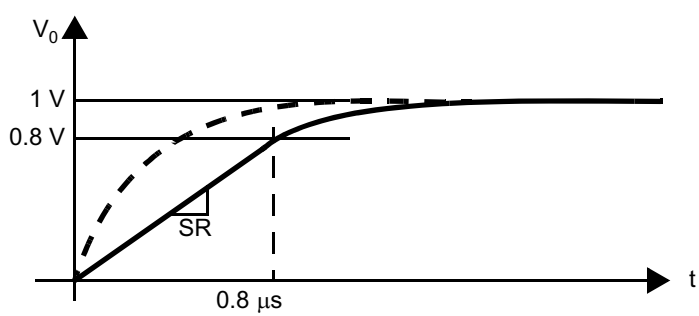


Fig. 6.7 The effect of slew rate limiting upon the settling of Example 6.4 with a 1-V step output.

7. To ensure this I_{D6} must be taken greater than I_{D5}, otherwise when the output voltage is slewing downward some current remains flowing in Q_2 and the total available current to charge C_C is only $I_{D6} < I_{D5}$.

Defining the slew rate to be the maximum rate that v_2 can change, and recalling that $v_{out} \cong v_2$, we have

$$SR \equiv \left.\frac{dv_{out}}{dt}\right|_{max} = \frac{I_{C_C}|_{max}}{C_C} = \frac{I_{D5}}{C_C} \tag{6.25}$$

where we used the charge equation $q = CV$, which leads to $I = dq/dt = C(dV/dt)$. Since $I_{D5} = 2I_{D1}$, we can also write

$$SR = \frac{2I_{D1}}{C_C} = \frac{I_{D5}}{C_C} \tag{6.26}$$

where I_{D1} is the original bias current of Q_1 with no signals present. Also, using (6.9), we have $C_C = g_{m1}/\omega_{ta}$, and substituting this into (6.26), we have

$$SR = \frac{2I_{D1}\omega_{ta}}{g_{m1}} \tag{6.27}$$

Recalling that

$$g_{m1} = \sqrt{2\mu_p C_{ox}\left(\frac{W}{L}\right)_1 I_{D1}} \tag{6.28}$$

and

$$V_{eff1} = \sqrt{\frac{2I_{D1}}{\mu_p C_{ox}(W/L)_1}} \tag{6.29}$$

from Chapter 1, we finally have another relationship for the slew-rate value.

$$SR = \frac{2I_{D1}}{\sqrt{2\mu_p C_{ox}(W/L)_1 I_{D1}}}\omega_{ta} = V_{eff1}\omega_{ta} \tag{6.30}$$

Since stability demands that ω_{ta} be lower than ω_{p2}, *the only ways of improving the slew rate for a properly-compensated two-stage CMOS opamp is to increase* V_{eff1} *or* ω_{p2}.

Key Point: *The only ways of improving the slew rate of a two-stage opamp is to increase* V_{eff1} *or* ω_{p2}.

EXAMPLE 6.5

Using the same parameters as in Example 6.1, and assuming $C_C = 1$ pF, what is the slew rate? What circuit changes could be made to double the slew rate but keep ω_{ta} and bias currents unchanged?

Solution

From (6.26), we have

$$SR = \frac{200\ \mu A}{1\ pF} = 200\ V/\mu s \tag{6.31}$$

To double the slew rate while maintaining the same bias currents, C_C should be set to 0.5 pF. To maintain the same unity-gain frequency, g_{m1} should be halved, which can be accomplished by decreasing the widths of Q_1 and Q_2 by 4 (i.e., one could let $W_1 = W_2 = 9\ \mu m$).

Assuming a fixed power consumption, and hence fixed bias currents, increasing V_{eff1} improves the slew rate (6.30) and helps to minimize distortion, but also lowers the transconductance of the input stage which decreases the dc gain (6.1), and increases the equivalent input thermal noise (see Chapter 9).

6.1.4 *n*-Channel or *p*-Channel Input Stage

The two-stage opamp shown in Fig. 6.3 and discussed thusfar has p-channel input transistors. It is also possible to realize a complementary opamp where the first stage has an n-channel differential pair and the second stage is a common-source amplifier having a p-channel input drive transistor. The choice of which configuration to use depends on a number of trade-offs that are discussed here.

First, the overall dc gain is largely unaffected by the choice since both designs have one stage with one or more n-channel driving transistors, and one stage with one or more p-channel driving transistors.

Having a p-channel input first stage implies that the second stage has an n-channel input drive transistor. This arrangement maximizes the transconductance of the drive transistor of the second stage, which is critical when high-frequency operation is important. As shown in Section 6.1.2, the equivalent second pole, and therefore the unity-gain frequency as well, are both proportional to the transconductance of the second stage.

Another consideration is whether a p-channel or n-channel source-follower output stage is desired. Typically, an n-channel source follower is preferable because this will have less of a voltage drop. Also, since an n-channel transistor has a higher transconductance, the effect on the equivalent second pole due to its load capacitance is minimized, which is another important consideration. Finally, there is less degradation of the gain when small load resistances are being driven. The one disadvantage of having an n-channel source follower is that, for n-well processes, it is not possible to connect the source to the substrate, thereby minimizing the voltage drop. Finally, note that, for opamps that drive purely capacitive loads, the buffer stage should not be included, and for this case, the output stage is clearly not a consideration.

Key Point: *When using a two-stage opamp, p-channel input transistors are almost always the best choice for the first stage because they offer larger unity-gain frequency and lower 1/f noise, with the major disadvantage being an increase in wideband thermal noise.*

Noise is another important consideration when choosing which input stage to use. Perhaps the major noise source of MOS opamps is due to 1/f noise caused by carriers randomly entering and leaving traps introduced by defects near the semiconductor surface. This 1/f noise source can be especially troublesome unless special circuit design techniques are used.[8] Typically, p-channel transistors have less 1/f noise than n-channel transistors since their majority carriers (holes) have less potential to be trapped in surface states. Thus, having a first-stage with p-channel inputs minimizes the output noise due to the 1/f noise. The same is not true when thermal noise is considered. When thermal noise is referred to the input of the opamp, it is minimized by using input transistors that have large transconductances, which unfortunately degrades the slew rate. However, when thermal noise is a major consideration, then a single-stage architecture, such as a folded-cascode opamp, is normally used.

In summary, when using a two-stage opamp, a p-channel input transistor for the first stage is almost always the best choice because it optimizes unity-gain frequency and minimizes 1/f noise, with the major disadvantage being an increase in wideband thermal noise.

6.1.5 Systematic Offset Voltage

When designing the two-stage CMOS opamp, if one is not careful, it is possible that the design will have an inherent (or systematic) input-offset voltage. Indeed, this was the case for many of the original designs used in

8. Some useful circuit techniques to reduce the effects of 1/f noise are correlated double sampling and chopper stabilization.

production integrated circuits. To see what is necessary to guarantee that no inherent input-offset voltage exists, consider the opamp shown in Fig. 6.3. When the differential input voltage is zero (i.e, when $V_{in}^{+} = V_{in}^{-}$), the output voltage of the first stage, V_{GS7}, should be that which is required to make I_{D7} equal to its bias current, I_{D6}. Specifically, the value of V_{GS7} should be given by

$$V_{GS7} = \sqrt{\frac{2I_{D6}}{\mu_n C_{ox}(W/L)_7}} + V_{tn} \tag{6.32}$$

When the differential input voltage is zero, the drain voltages of both Q_3 and Q_4 are equal by arguments of symmetry. Therefore, the output voltage of the first stage, V_{GS7}, is given by

$$V_{GS7} = V_{DS3} = V_{GS4} \tag{6.33}$$

This value is the voltage necessary to cause I_{D7} to be equal to I_{D6}, or, in other words, to satisfy (6.32). If this value is not achieved, then the output of the second stage (with Q_6, Q_7) would clip at either the negative or positive rail since this stage has such a high gain.[9] However, the gate-source voltage of Q_4 is given by

$$V_{GS4} = \sqrt{\frac{2I_{D4}}{\mu_n C_{ox}(W/L)_4}} + V_{tn} \tag{6.34}$$

so equating (6.32) and (6.34) to satisfy (6.33) results in

$$\sqrt{\frac{2I_{D4}}{\mu_n C_{ox}(W/L)_4}} = \sqrt{\frac{2I_{D6}}{\mu_n C_{ox}(W/L)_7}} \tag{6.35}$$

or, equivalently,

$$\frac{I_{D4}}{(W/L)_4} = \frac{I_{D6}}{(W/L)_7} \tag{6.36}$$

This equality, when the current density of Q_4 is equal to the current density of Q_7, guarantees that they both have the same effective gate-source voltages.

Since

$$\frac{I_{D6}}{I_{D4}} = \frac{I_{D6}}{I_{D5}/2} = \frac{(W/L)_6}{(W/L)_5/2} \tag{6.37}$$

we see that the necessary condition to ensure that no input-offset voltage is present is

$$\frac{(W/L)_7}{(W/L)_4} = 2\frac{(W/L)_6}{(W/L)_5} \tag{6.38}$$

which is satisfied for the sizes shown. Note that this analysis ignores any voltage drop in a output buffer source-follower stage (not shown in Fig. 6.3) and any mismatches between the output impedance of p-channel and n-channel transistors. Fortunately, these effects cause only minor offset voltages, and by satisfying (6.38), offset voltages on the order of 5 mV or less can be obtained.

9. When the opamp is incorrectly designed to have an inherent input-offset voltage, feedback will cause the differential input voltage of the first stage to be equal to a nonzero voltage needed to ensure that the second stage does not clip.

EXAMPLE 6.6

Consider the opamp of Fig. 6.3, where Q_3 and Q_4 are each changed to widths of 10 μm, and we want the output stage to have a bias current of 400 μA. Find the new sizes of Q_6 and Q_7, such that there is no systematic offset voltage.

Solution

Since I_{D6} determines the bias current of the output stage, and since it should have 2-times more current than I_{D5}, its width should be 2-times greater than W_5, resulting in

$$W_6 = 60\ \mu m \tag{6.39}$$

For Q_7, we use (6.38), which leads to

$$W_7 = \frac{2(60)(10)}{30}\mu m = 40\ \mu m \tag{6.40}$$

6.2 OPAMP COMPENSATION

This section discusses using opamps in closed-loop configurations and how to compensate an opamp to ensure that the closed-loop configuration is not only stable, but also has good settling characteristics. Although the two-stage opamp is used as an example, almost all the material discussed here applies to other opamps as well.

Optimum compensation of opamps is typically considered to be one of the most difficult parts of the opamp design procedure. However, if the systematic approach taken here is used, then a straightforward procedure can be followed that almost always results in a near-optimum compensation network.

6.2.1 Dominant-Pole Compensation and Lead Compensation

The two most popular tools available to analog circuit designers for opamp compensation are dominant-pole compensation and lead compensation.

Key Point: *In general, dominant-pole compensation is performed by decreasing the frequency of one pole in a feedback loop while leaving the others relatively unchanged.*

Dominant-pole compensation is performed by forcing a feedback system's open-loop response to be closely approximated by a first-order response up to the loop's unity gain frequency. As described in Section 5.3.1, first-order feedback systems are unconditionally stable with at least 90° phase margin. Unfortunately, opamp circuits generally have multiple poles and zeros. Increasing the frequency of poles in a circuit is not generally practical since that would demand increased power consumption or some other undesirable trade-off. Hence, the easiest way to make the system behave like a first-order system is to intentionally decrease the frequency of one dominant pole, making it much lower than the other poles and zeros of the system. The idea is illustrated in Fig. 6.8 which plots the open loop response, $L(\omega)$. Recall that $L(\omega)$ is the product of the amplifier response and the feedback network response, $A(\omega)\beta$. During dominant-pole compensation the pole frequency ω_{p1} has been decreased to a new much lower frequency. The result is a decrease in the unity-gain frequency of L, ω_t, and an attendant increase in phase margin.

A further increase in phase margin is obtained by lead compensation which introduces a left half-plane zero, ω_z, at a frequency slightly greater than ω_t. If done properly, this has minimal effect on the unity-gain frequency, but does introduce an additional approximately +20° phase shift around the unity-gain frequency. This results in a higher unity-gain frequency for a specified phase margin than would be attainable using dominant-pole compensation alone.

In Fig. 6.8, both dominant-pole and lead compensation are illustrated having minimal impact on the dc loop gain and the other pole and zero frequencies. Depending upon the specific circuit implementation of the compensation, the situation may be somewhat more complicated. The approach is next described in detail for a classic two-stage opamp.

Key Point: *Lead compensation in opamp circuits introduces into the loop response a left half-planezero at a frequency slightly greater than the unity gain frequency to provide a small phase lead and, therefore, and increase in phase margin.*

6.2.2 Compensating the Two-Stage Opamp

Stability and phase margin depend upon the loop gain of a feedback amplifier, $L(s)$, whereas our analysis so far has focused on the response of the amplifier alone, $A_v(s)$. *The feedback network must be accounted for in compensation.* It was shown in Section 5.4 that for either inverting or noninverting opamp configurations, the undriven circuit is as shown in Fig. 6.9 and that assuming the opamp has relatively high input impedance and an output impedance smaller than the load or feedback network, the loop gain may be approximated as

Key Point: *The feedback network must be accounted for in compensation.*

$$L(s) \approx A_v(s)\frac{Z_1}{Z_1 + Z_2} \tag{6.41}$$

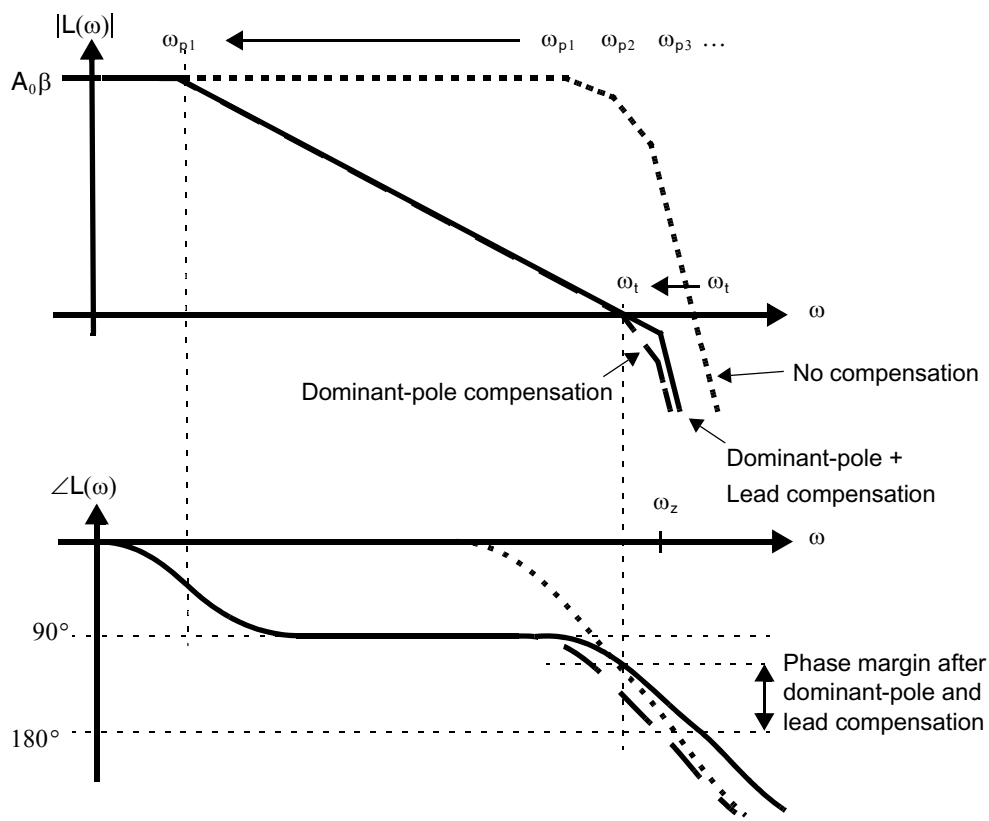


Fig. 6.8 A Bode plot of open-loop gain illustrating dominant-pole and lead compensation.

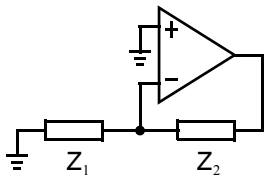


Fig. 6.9 An undriven opamp circuit with a general feedback network.

Although there are several different possible ways to factor $L(s)$ into $A(s)$ and β,[10] the following straightforward approximation $\beta = Z_1/(Z_1 + Z_2)$ may be adopted and it is assumed for analysis that β is constant up to the loop's unity gain frequency. For example, in a unity gain configuration, $\beta = 1$ and $L(s) \cong A_v(s)$. For an accurate determination of phase margin, simulation may be used to properly account for the impact of loading.

The capacitor, C_C, controls the dominant first pole, (i.e., ω_{p1}), and thereby the loop's unity-gain frequency, ω_t.

$$\omega_t = L_0\omega_{p1} = \beta g_{m1}/C_C \tag{6.42}$$

Hence, by properly selecting the value of C_C dominant-pole compensation can be achieved.

Lead compensation is achieved using R_C. If the small-signal model of Fig. 6.5 is reanalyzed with a nonzero R_C, then a third-order denominator results. The first two poles are still approximately at the frequencies given by (6.19) and (6.20). The third pole is at a high frequency and has almost no effect. However, the zero is now determined by the relationship

$$\omega_z = \frac{-1}{C_C(1/g_{m7} - R_C)} \tag{6.43}$$

This result allows the designer a number of possibilities. One could take

$$R_C = 1/g_{m7} \tag{6.44}$$

to eliminate the right-half-plane zero altogether. Alternatively, one could choose R_C to be larger and thus move the right-half-plane zero into the left half plane to cancel the nondominant pole, ω_{p2}. Setting (6.43) equal to (6.20) and solving for R_C results in the following equation for R_C:

$$R_C = \frac{1}{g_{m7}}\left(1 + \frac{C_1 + C_2}{C_C}\right) \tag{6.45}$$

Unfortunately, C_2 is often not known a priori, especially when no output stage is present.

The third possibility (recommended by the authors) is to choose R_C even larger to move the now left-half-plane zero to a frequency slightly greater than the unity-gain frequency that would result if the lead resistor were not present [Roberge, 1975]. For example, if the new zero frequency is taken 70% higher than the unity-gain frequency, it will introduce a phase lead of $\tan^{-1}(1/1.7) = 30°$. For this case, one should satisfy the following equation:

$$\omega_z = 1.7\omega_t \tag{6.46}$$

Assuming $R_C >> (1/g_{m7})$, then $\omega_z \cong 1/(R_C C_C)$. Recalling from (6.42) that $\omega_t = \beta g_{m1}/C_C$, then one should choose R_C according to

$$R_C \cong \frac{1}{1.7\beta g_{m1}} \tag{6.47}$$

10. In Chapter 5, β was defined as the inverse of the desired closed-loop gain. Therefore, depending upon the configuration and application of the circuit, β may be taken as $-Z_1/Z_2$ (inverting configuration), $Z_1/(Z_1 + Z_2)$ (noninverting configuration), or $(-1)/Z_2$ (transimpedance feedback amplifier).

Finally, the lead compensation resistor R_C may be replaced by a transistor operating in the triode region, as illustrated in Fig. 6.10. Transistor Q_9 has $V_{DS9} = 0$ since no dc bias current flows through it, and therefore Q_9 is deep in the triode region. Thus, this transistor operates as a resistor, R_C, with a value given by

$$R_C = r_{ds9} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_9 V_{eff9}} \tag{6.48}$$

It should be noted here that r_{ds} indicates the drain-source resistance of Q_9 when it is in the triode region as opposed to the finite-output impedance of Q_9 when it is in the active mode. The same notation, r_{ds}, is used to indicate the drain-source resistance in both cases—whether the transistor is in the active or the triode region. One simply has to check which region a transistor is operating in to ensure that the correct equation is used to determine r_{ds}.

This approach leads to the following design procedure for compensation of a two-stage CMOS opamp:

Key Point: *Compensation of the two-stage opamp proceeds by selecting Miller capacitor C_C to provide 55 degrees phase margin, then introducing a lead compensation series resistor to add another 30 degrees of phase margin.*

1. Start by choosing, somewhat arbitrarily, $C_C' \cong (\beta g_{m1}/g_{m7})C_L$. This initially places the loop's unity gain frequency (6.42) approximately at the frequency of the second pole (6.20), where it has been assumed that the load capacitance C_L is dominant.
2. Using SPICE, find the frequency at which a –125° phase shift exists. Let the gain at this frequency be denoted A'. Also, let the frequency be denoted ω_t. This is the frequency that we would like to become the unity-gain frequency of the loop gain.
3. Choose a new C_C so that ω_t becomes the unity-gain frequency of the loop gain, thus resulting in a 55° phase margin. (Obtaining this phase margin is the reason we chose –125° in step 2.) This can be achieved by taking C_C according to the equation

$$C_C = C_C' A' \tag{6.49}$$

   It might be necessary to iterate on C_C a couple of times using SPICE.
4. Choose R_C according to

$$R_C = \frac{1}{1.7\omega_t C_C} \tag{6.50}$$

   This choice will increase the phase margin approximately 30° resulting in a total phase margin of approximately 85°. It allows a margin of 5° to account for processing variations without the poles of the closed-loop response becoming real. *This choice is also almost optimum lead compensation for almost any opamp*

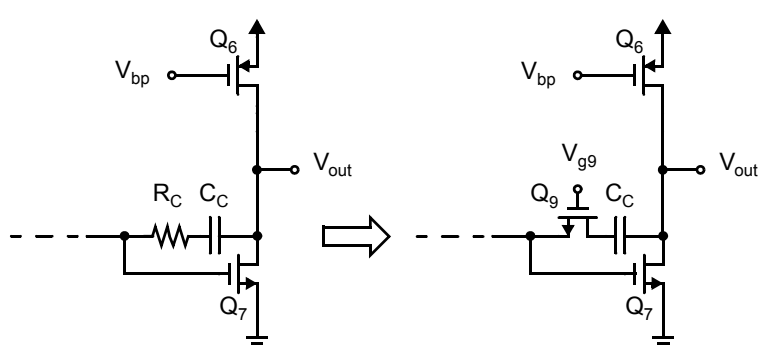


Fig. 6.10 The common-source second stage of a CMOS two-stage amplifier, showing the lead compensation resistor R_C replaced by triode transistor Q_9.

when a resistor is placed in series with the compensation capacitor. It might be necessary to iterate on R_C a couple of times to optimize the phase margin because the unity-gain frequency may be shifted somewhat by the lead compensation, and because (6.50) assumes $R_C >> (1/g_{m7})$. However, one should check that the gain continues to steadily decrease at frequencies above the new unity-gain frequency; otherwise, the transient response can be poor. This situation sometimes occurs when unexpected zeros at frequencies only slightly greater than ω_t are present.

5. If, after step 4, the phase margin is not adequate, then increase C_C while leaving R_C constant. This will move both ω_t and the lead zero to lower frequencies while keeping their ratio approximately constant, thus minimizing the effects of higher-frequency poles and zeros which, hopefully, do not also move to lower frequencies. In most cases, the higher-frequency poles and zeros (except for the lead zero) will not move to significantly lower frequencies when C_C is increased.
6. The final step is to replace R_C by a transistor. The size of the transistor can be chosen using equation (6.48), which is repeated here for convenience:

$$R_C = r_{ds9} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_9 V_{eff9}} \tag{6.51}$$

Finally, SPICE can be used again to fine-tune the device dimensions to optimize the phase margin to that obtained in steps 4 and 5.

Not only does the procedure just described apply to the two-stage opamp, but it (or a very similar procedure) has been found to be almost optimum for compensating most types of opamps.

EXAMPLE 6.7

An opamp has an open-loop transfer function given by

$$A(s) = \frac{A_0(1 + s/\omega_z)}{(1 + s/\omega_{p1})(1 + s/\omega_2)} \tag{6.52}$$

Here, A_0 is the dc gain of the opamp and ω_z, ω_p, and ω_2 are the frequencies of a zero, the dominant pole, and the equivalent second pole, respectively. Assume that $\omega_2 = 2\pi \times 50$ MHz and that $A_0 = 10^4$. The opamp is to be used in a unity-gain configuration so that $\beta = 1$ and $L(s) = A(s)$.

a. Assuming $\omega_z \to \infty$, find ω_{p1} and the unity-gain frequency, ω_t', so that the opamp has a unity-gain phase margin of 55°.

b. Assuming $\omega_z = 1.7\omega_t'$ (where ω_t' is as found in part (a)), what is the new unity-gain frequency, ω_t? Also, find the new phase margin.

Solution

First, note that at frequencies much greater than the dominant-pole frequency (i.e., $\omega >> \omega_{p1}$), such as the unity-gain frequency, (6.52) can be approximated by

$$L(s) = A(s) \cong \frac{A_0(1 + s/\omega_z)}{(s/\omega_{p1})(1 + s/\omega_2)} \tag{6.53}$$

(*a*) For $\omega_z \to \infty$ we use (6.53) to find the phase angle at ω_t':

$$\angle A(j\omega_t') = -90° - \tan^{-1}(\omega_t'/\omega_2) \tag{6.54}$$

For a phase margin of $\mathrm{PM} = 55°$, we need

$$\angle A(j\omega_t') = -180° + \mathrm{PM} = -125° \tag{6.55}$$

Setting (6.54) equal to (6.55) and solving for ω_t'/ω_2 results in

$$\begin{aligned} \tan^{-1}(\omega_t'/\omega_2) &= 35° \\ \Rightarrow \omega_t' &= 2.2 \times 10^8 \text{rad/s} = 2\pi \times 35 \text{ MHz} \end{aligned} \tag{6.56}$$

Next, setting $|A(j\omega_t')| = 1$ and again using (6.53) results in

$$\begin{aligned} \frac{A_0}{(\omega_t'/\omega_{p1})\sqrt{1 + (\omega_t'/\omega_2)^2}} &= 1 \\ \Rightarrow \omega_{p1} = \frac{\omega_t'\sqrt{1 + (\omega_t'/\omega_2)^2}}{A_0} &= 2\pi \times 4.28 \text{ kHz} \end{aligned} \tag{6.57}$$

(*b*) First, we set

$$\omega_z = 1.7\omega_t' = 2\pi \times 59.5 \text{ MHz} \tag{6.58}$$

To find the new unity-gain frequency, setting $|A(j\omega_t)| = 1$ in (6.53) now gives

$$\begin{aligned} \frac{A_0\sqrt{1 + (\omega_t/\omega_z)^2}}{(\omega_t/\omega_{p1})\sqrt{1 + (\omega_t/\omega_2)^2}} &= 1 \\ \Rightarrow \omega_t = \frac{A_0\omega_{p1}\sqrt{1 + (\omega_t/\omega_z)^2}}{\sqrt{1 + (\omega_t/\omega_2)^2}} & \end{aligned} \tag{6.59}$$

This equation can be solved for the new unity-gain frequency. Both sides can be squared, resulting in a quadratic equation in ω_t, which can be solved exactly. Alternatively, the value for ω_t' found in part (*a*) can be used as an initial guess, and (6.59) can be solved iteratively. After four iterations, one finds that $\omega_t = 2\pi \times 39.8$ MHz. This unity-gain frequency is a 14 percent increase over that found in part (*a*). We can solve for the new phase shift by using

$$\angle A(j\omega_t) = -90° + \tan^{-1}(\omega_t/\omega_z) - \tan^{-1}(\omega_t/\omega_2) = -95° \tag{6.60}$$

Note that this phase value gives a phase margin of 85°, which is a 30° improvement. In practice, the improvement may not be this great due to additional high-frequency poles and zeros (which have been ignored here). Regardless of this degradation, the improvement from using lead compensation is substantial.

6.2.3 Making Compensation Independent of Process and Temperature

This section shows how lead compensation can be made process and temperature insensitive. Repeating equations (6.9) and (6.20) here, we have

$$\omega_t = \frac{g_{m1}}{C_C} \tag{6.61}$$

and

$$\omega_{p2} \cong \frac{g_{m7}}{C_1 + C_2} \tag{6.62}$$

We see here that the second pole is proportional to the transconductance of the drive transistor of the second stage, g_{m7}. Also, the unity-gain frequency is proportional to the transconductance of the input transistor of the first stage, g_{m1}. Furthermore, the ratios of all of the transconductances remain relatively constant over process and temperature variations since the transconductances are all determined by the same biasing network.[11] Also, most of the capacitances also track each other since they are primarily determined by gate oxides. Repeating (6.43), when a resistor is used to realize lead compensation, the lead zero is at a frequency given by

$$\omega_z = \frac{-1}{C_C(1/g_{m7} - R_C)} \tag{6.63}$$

Thus, if R_C can also be made to track the inverse of transconductances, and in particular $1/g_{m7}$, then the lead zero will also be proportional to the transconductance of Q_7. As a result, the lead zero will remain at the same relative frequency with respect to ω_t and ω_{p2}, as well as all other high-frequency poles and zeros. In other words, the lead compensation will be mostly independent of process and temperature variations.

It turns out that R_C can be made proportional to $1/g_{m7}$ as long as R_C is realized by a transistor in the triode region that has an effective gate-source voltage proportional to that of Q_7. To see this result, recall that R_C is actually realized by Q_9, and therefore we have

$$R_C = r_{ds9} = \frac{1}{\mu_n C_{ox}(W/L)_9 V_{eff9}} \tag{6.64}$$

Also, g_{m7} is given by

$$g_{m7} = \mu_n C_{ox}(W/L)_7 V_{eff7} \tag{6.65}$$

Thus, the product $R_C g_{m7}$, which we want to be a constant, is given by

$$R_C g_{m7} = \frac{(W/L)_7 V_{eff7}}{(W/L)_9 V_{eff9}} \tag{6.66}$$

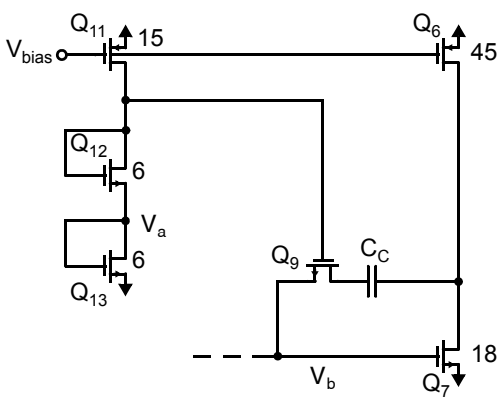


Fig. 6.11 The bias circuit, second stage, and compensation circuit of the two-stage opamp.

Therefore, all that remains is to ensure that V_{eff9}/V_{eff7} is independent of process and temperature variations since clearly the remaining terms depend only on a geometric relationship. The ratio V_{eff9}/V_{eff7} can be made constant by deriving V_{GS9} from the same biasing circuit used to derive V_{GS7}. Specifically, consider the circuit shown in Fig. 6.11 which consists of a bias stage, the second stage, and the compensation network of the two-stage opamp. First we must make $V_a = V_b$, which is possible by making $V_{eff13} = V_{eff7}$. These two effective gate-source voltages can be made equal by taking

$$\sqrt{\frac{2I_{D7}}{\mu_n C_{ox}(W/L)_7}} = \sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}} \tag{6.67}$$

Squaring and simplifying, we see that the following condition must be satisfied:

$$\frac{I_{D7}}{I_{D13}} = \frac{(W/L)_7}{(W/L)_{13}} \tag{6.68}$$

11. The ratio μ_n/μ_p' is relatively constant for a given process but often varies considerably from process to process (say, from 0.6-μm to 0.4-μm CMOS processes).

However, the ratio I_{D7}/I_{D13} is set from the current mirror pair Q_6, Q_{11}, resulting in

$$\frac{I_{D7}}{I_{D13}} = \frac{(W/L)_6}{(W/L)_{11}} \tag{6.69}$$

Thus, to make $V_{eff13} = V_{eff7}$, from (6.68) and (6.69), we need to satisfy the following relationship:

$$\frac{(W/L)_6}{(W/L)_7} = \frac{(W/L)_{11}}{(W/L)_{13}} \tag{6.70}$$

Assuming this condition is satisfied, we have $V_{eff13} = V_{eff7}$, and therefore we also have $V_a = V_b$. Since the gates of Q_{12} and Q_9 are connected and their source voltages are the same, we also have $V_{eff12} = V_{eff9}$. Next, using these gate-source relationships and noting that $I_{D12} = I_{D13}$, we can write

$$\frac{V_{eff7}}{V_{eff16}} = \frac{V_{eff13}}{V_{eff12}} = \frac{\sqrt{\dfrac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}}}{\sqrt{\dfrac{2I_{D12}}{\mu_n C_{ox}(W/L)_{12}}}} = \sqrt{\frac{(W/L)_{12}}{(W/L)_{13}}} \tag{6.71}$$

Finally, substituting (6.71) into (6.66), we have the product $R_C g_{m7}$, given by

$$R_C g_{m7} = \frac{(W/L)_7}{(W/L)_9}\sqrt{\frac{(W/L)_{12}}{(W/L)_{13}}} \tag{6.72}$$

which is only dependent on geometry and not on processing or temperature variations.

As a result, we have guaranteed that the drain-source resistance of a transistor in the triode region is inversely matched to the transconductance of a different transistor. This relationship can be very useful for many other applications as well. Indeed, in Chapter 7 we will see that it's quite simple to make all of the transconductances of transistors in a microcircuit match the conductance of a single off-chip resistor. *This approach results in the possibility of on-chip "resistors," realized by using triode-region transistors that are accurately ratioed with respect to a single off-chip resistor.* This relationship can be very useful in modern circuit design.

6.3 ADVANCED CURRENT MIRRORS

6.3.1 Wide-Swing Current Mirrors

As newer technologies with shorter channel lengths are used, it becomes more difficult to achieve reasonable opamp gains due to transistor output-impedance degradation caused by short-channel effects. As a result, designers are often forced to use cascode current mirrors. Unfortunately, the use of conventional cascode current mirrors limits the signal swings available, which may not be tolerated in certain applications. Fortunately, circuits exist that do not limit the signal swings as much as the current mirrors discussed in Chapter 3. One such circuit is shown in Fig. 6.12 and is often called the "wide-swing cascode current mirror" [Sooch, 1985; Babanezhad, 1987].

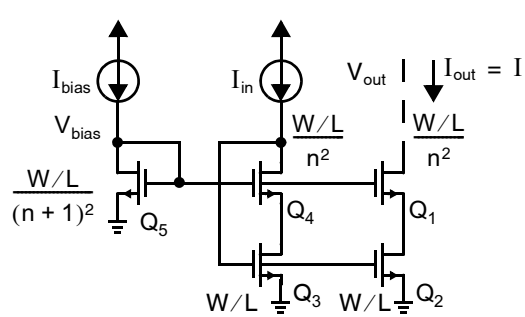


Fig. 6.12 The wide-swing cascode current.

The basic idea of this current mirror is to bias the drain-source voltages of transistors Q_2 and Q_3 to be close to the minimum possible without them going into the triode region. Specifically, if the sizes shown in Fig. 6.12 are used, and assuming the classical square-law equations for long channel-length devices are valid, transistors Q_2 and Q_3 will be biased right at the edge of the triode region. Before seeing how these bias voltages are created, note that the transistor pair Q_3, Q_4 acts like a single diode-connected transistor in creating the gate-source voltage for Q_3. These two transistors operate very similarly to how Q_3 alone would operate if its gate were connected to its drain. The reason for including Q_4 is to lower the drain-source voltage of Q_3 so that it is matched to the drain-source voltage of Q_2. This matching makes the output current, I_{out}, more accurately match the input current, I_{in}. If Q_4 were not included, then the output current would be a little smaller than the input current due to the finite output impedances of Q_2 and Q_3. Other than this, Q_4 has little effect on the circuit's operation.

To determine the bias voltages for this circuit, let V_{eff} be the effective gate-source voltage of Q_2 and Q_3, and assume all of the drain currents are equal. We therefore have

$$V_{eff} = V_{eff2} = V_{eff3} = \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}(W/L)}} \tag{6.73}$$

Furthermore, since Q_5 has the same drain current but is $(n+1)^2$ times smaller, we have

$$V_{eff5} = (n+1)V_{eff} \tag{6.74}$$

Similar reasoning results in the effective gate-source voltages of Q_1 and Q_4 being given by

$$V_{eff1} = V_{eff4} = nV_{eff} \tag{6.75}$$

Thus,

$$V_{G5} = V_{G4} = V_{G1} = (n+1)V_{eff} + V_{tn} \tag{6.76}$$

Furthermore,

$$V_{DS2} = V_{DS3} = V_{G5} - V_{GS1} = V_{G5} - (nV_{eff} + V_{tn}) = V_{eff} \tag{6.77}$$

This drain-source voltage puts both Q_2 and Q_3 right at the edge of the triode region. Thus, the minimum allowable output voltage is now

$$V_{out} > V_{eff1} + V_{eff2} = (n+1)V_{eff} \tag{6.78}$$

A common choice for n might be simply unity, in which case the current mirror operates correctly as long as

$$V_{out} > 2V_{eff} \tag{6.79}$$

With a typical value of V_{eff} between 0.2 V and 0.25 V, the wide-swing current mirror can guarantee that all of the transistors are in the active (i.e., saturation) region even when the voltage drop across the mirror is as small as 0.4 V to 0.5 V.

However, there is one other requirement that must be met to ensure that all transistors are in the active region. Specifically, we need

$$V_{DS4} > V_{eff4} = nV_{eff} \tag{6.80}$$

to guarantee that Q_4 is in the active region. To find V_{DS4}, we note that the gate of Q_3 is connected to the drain of Q_4, resulting in

$$V_{DS4} = V_{G3} - V_{DS3} = (V_{eff} + V_{tn}) - V_{eff} = V_{tn} \tag{6.81}$$

As a result, one need only ensure that V_{tn} be greater than nV_{eff} for Q_4 to remain in the active region—not a difficult requirement.

Key Point: *Cascoding is a key technique to enable high gain using transistors with low drain-source resistance, but if the cascode transistor is improperly biased, output swing is greatly reduced.*

It should be noted that this circuit was analyzed assuming the bias current, I_{bias}, equals the input current, I_{in}. Since, in general, I_{in} may be a varying current level, there is some choice as to what I_{bias} value should be chosen. One choice is to set I_{bias} to the largest expected value for I_{in}. Such a choice will ensure that none of the devices exit their active region, though

the drain-source voltage of Q_2 and Q_3 will be larger than necessary except when the maximum I_{in} is applied. As a result, some voltage swing will be lost. Perhaps the more common choice in a wide-swing opamp is to let I_{bias} equal the nominal value of I_{in}. With this setting, some devices will enter triode and the output impedance will be reduced for larger I_{in} values (say, during slew-rate limiting), but such an effect during transient conditions can often be tolerated.

Before leaving this circuit, a few design comments are worth mentioning. In most applications, an experienced designer would take $(W/L)_5$ smaller than the size given in Fig. 6.12 to bias transistors Q_2 and Q_3 with slightly larger drain-source voltages than the minimum required (perhaps 0.1 V larger). This increase accounts for the fact that practical transistors don't have a sharp transition between the triode and active regions. This increase would also help offset a second-order effect due to the body effect of Q_1 and Q_4, which causes their threshold voltages to increase, which tends to push Q_2 and Q_3 more into the triode region. In addition, to save on power dissipation, the bias branch consisting of Q_5 and I_{bias} might be scaled to have lower currents while keeping the same current densities and, therefore, the same effective gate-source voltages. A final common modification is to choose the lengths of Q_2 and Q_3 just a little larger than the minimum allowable gate length (as the drain-source voltage across them is quite small), but Q_1 and Q_4 would be chosen to have longer gate lengths since the output transistor (i.e., Q_1) often has larger voltages across it. A typical size might be twice the minimum allowable channel length. This choice of gate lengths helps eliminate detrimental short-channel effects, such as the drain-to-substrate leakage currents of Q_1 that might otherwise result. Minimizing the lengths of Q_2 and Q_3 maximizes the frequency response, as their gate-source capacitances are the most significant capacitances contributing to high-frequency poles.

6.3.2 Enhanced Output-Impedance Current Mirrors and Gain Boosting

Another variation on the cascode current mirror is often referred to as the *enhanced output-impedance current mirror*. A simplified form of this circuit is shown in Fig. 6.13, and it is used to increase the output impedance. The basic idea is to use a feedback amplifier to keep the drain-source voltage across Q_2 as stable as possible, irrespective of the output voltage. The addition of this amplifier ideally increases the output impedance by a factor equal to one plus the loop gain over that which would occur for a classical cascode current mirror. Specifically, using the results from Chapter 3,

$$R_{out} \cong g_{m1} r_{ds1} r_{ds2} (1 + A) \tag{6.82}$$

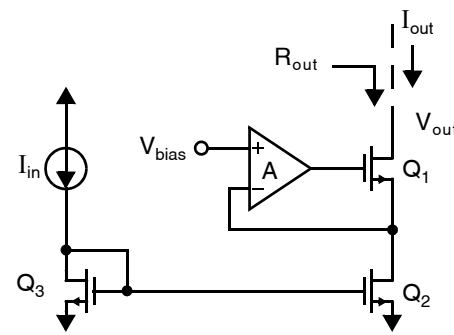


Fig. 6.13 The enhanced output-impedance current mirror.

In practice, the output impedance might be limited by a parasitic conductance between the drain of Q_1 and its substrate due to short-channel effects. This parasitic conductance is a result of collisions between highly energized electrons resulting in electron-hole pairs to be generated with the holes escaping to the substrate. The generation of these electron-hole pairs is commonly called *impact ionization*. Stability of the feedback loop comprising the amplifier $A(s)$ and Q_1 must of course be verified.

The same technique can be applied to increase the output impedance of a cascode gain stage, as shown in Fig. 6.14. The gain of the stage is given by

$$A_V(s) = \frac{V_{out}(s)}{V_{in}(s)} = -g_{m2}\left(R_{out}(s) \parallel \frac{1}{sC_L}\right) \tag{6.83}$$

where R_{out} is now a function of frequency and given by

$$R_{out}(s) = g_{m1} r_{ds1} r_{ds2} (1 + A(s)) \tag{6.84}$$

Key Point: *Auxiliary amplifiers may be used to enhance the output impedance of cascode current mirrors and gain stages, thereby boosting dc gain.*

Note that the gain is increased over a conventional cascode amplifier by a factor $(1 + A(s))$. The technique is therefore called *gain-boosting*, and may be used to improve the dc gain of opamp circuits described later in this chapter. Note that gain-boosting is only realized if the output impedance the current mirror active load, denoted I_{bias} in Fig. 6.14, is similarly enhanced.

EXAMPLE 6.8

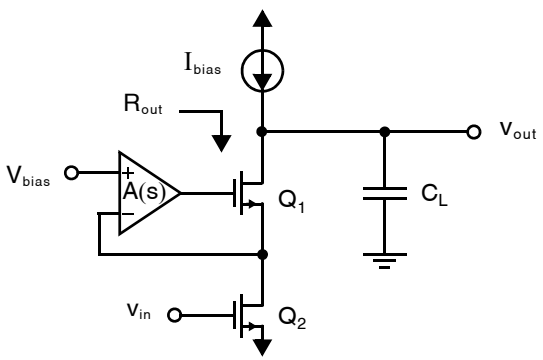


Fig. 6.14 A cascode amplifier with gain-boosting.

Assume an amplifier, $A(s)$, has a transfer function given by

$$A(s) = \frac{A_0}{1 + s\tau'_1} \cong \frac{A_0}{s\tau'_1} \tag{6.85}$$

around its unity-gain frequency. Find approximate equations for any additional poles and/or zeros introduced by the gain enhancement circuit that are significant around the unity-gain frequency of the common-source amplifier shown in Fig. 6.14.

Solution

Note that at high frequencies, it is not possible to assume that $|A(j\omega)| >> 1$. Working with conductances, we have

$$G_{out}(s) = \frac{1}{R_{out}(s)} = \frac{g_{ds1}g_{ds2}}{g_{m1}(1 + A(s))} \tag{6.86}$$

which can be inserted into (6.83) giving

$$\begin{aligned} A_V(s) &= \frac{-g_{m2}}{sC_L + G_{out}(s)} = \frac{-g_{m2}}{sC_L + \dfrac{g_{ds1}g_{ds2}}{g_{m1}(1 + A(s))}} \\ &= \frac{-g_{m2}(1 + A(s))}{sC_L(1 + A(s)) + \dfrac{g_{ds1}g_{ds2}}{g_{m1}}} \end{aligned} \tag{6.87}$$

Substituting (6.85) into (6.87) and rearranging gives

$$A_V(s) = \frac{-g_{m2}(A_0 + s\tau'_1)}{s\left(A_0C_L + \dfrac{g_{ds1}g_{ds2}\tau'_1}{g_{m1}} + sC_L\tau'_1\right)} \tag{6.88}$$

From (6.88), we see that there is a left-hand-plane zero at a frequency given by

$$\omega_z = \frac{A_0}{\tau'_1} \tag{6.89}$$

which is the approximate unity-gain frequency of the amplifier, $A(s)$. Also, there are two poles, but the pole at dc is not accurate since we have made use of approximate equations that are valid only near the unity-gain frequency

of the amplifier. However, the other pole location is a good approximation and it occurs in the left-half-plane given by (after some rearranging)

$$p_2 = \frac{A_0}{\tau_1'} + \frac{g_{ds1}g_{ds2}}{C_L g_{m1}} = \frac{A_0}{\tau_1'} + \frac{1}{C_L R'} \tag{6.90}$$

Here, we have defined

$$R' = g_{m1} r_{ds1} r_{ds2} \tag{6.91}$$

which is approximately the output impedance of the cascode mirror without enhancement and is quite a large value. Furthermore, note that the term A_0/τ_1' is a value near the unity-gain frequency of the amplifier, and therefore, typically the first term in (6.90) dominates. Thus, the added pole frequency is approximately given by

$$p_2 \cong \frac{A_0}{\tau_1'} \cong \omega_z' \tag{6.92}$$

Therefore, like the new zero, the new pole also occurs around the unity-gain frequency of the amplifier, and their effects mostly cancel each other out. In addition, if the unity-gain frequency of the enhancement loop is greater than the unity-gain frequency of the opamp, the frequency response of the enhancement amplifier shouldn't have a large detrimental effect on the overall frequency response.

The enhanced output-impedance current mirror appears to have been originally proposed in [Hosticka, 1979] for single-ended input amplifiers and more recently described almost simultaneously in [Bult, 1990] and [Säckinger, 1990] for differential-input amplifiers. Bult proposed that complete opamps be used for each current mirror. In a modern fully differential opamp, this might increase the number of opamps by four, although these extra opamps can be scaled to have less power dissipation. The implementation proposed by Säckinger is shown in Fig. 6.15. The feedback amplifier in this case is realized by the common-source amplifier consisting of Q_3 and its current source I_{B1}. Assuming the output impedance of current source I_{B1} is approximately equal to r_{ds3}, the loop gain will be $(g_{m3}r_{ds3})/2$, and the final ideal output impedance will be given by

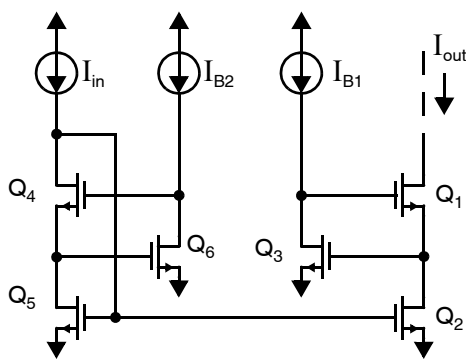


Fig. 6.15 The Säckinger implementation of the enhanced output-impedance current mirror.

$$r_{out} \cong \frac{g_{m1}g_{m3}r_{ds1}r_{ds2}r_{ds3}}{2} \tag{6.93}$$

The circuit consisting of Q_4, Q_5, Q_6, I_{in}, and I_{B2} operates almost identically to a diode-connected transistor, but is used instead to guarantee that all transistor bias voltages are accurately matched to those of the output circuitry consisting of Q_1, Q_2, Q_3, and I_{B1}. As a result, I_{out} will very accurately match I_{in}.

The Säckinger realization is much simpler than that proposed by Bult, but has a major limitation in that the signal swing is significantly reduced. This reduction is a result of Q_2 and Q_5 being biased to have drain-source voltages much larger than the minimum required. Specifically, their drain-source voltages are given by

$$V_{DS2} = V_{DS5} = V_{eff3} + V_{tn} \tag{6.94}$$

rather than the minimum required, which would be equal to V_{eff2}. This limitation is especially harmful for modern technologies operating with power supply voltages of 3.3 V or lower. In the next subsection, an alternative realization is described that combines the wide-swing current mirror with the enhanced output-impedance circuit.

6.3.3 Wide-Swing Current Mirror with Enhanced Output Impedance

This current mirror was originally described in [Gatti, 1990] for use in current-mode continuous-time filters and was then used in the design of wide-signal-swing opamps [Coban, 1994; Martin, 1994]. It is very similar to the enhanced output-impedance current mirrors of [Säckinger, 1990], except that diode-connected transistors used as level shifters have been added in front of the common-source enhancement amplifiers, as shown in Fig. 6.16. At the output side, the level shifter is the diode-connected transistor, Q_4, biased with current I_{bias}. The circuitry at the input basically acts as a diode-connected transistor while ensuring that all bias voltages are matched to the output circuitry so that I_{out} accurately matches I_{in}. Shown next to each transistor is a reasonable width in μm. Note that for the case in which

$$I_{bias} = I_{in}/7 \tag{6.95}$$

all transistors are biased with nearly the same current density, except for Q_3 and Q_7. As a result, all transistors have the same effective gate-source voltages, V_{eff}, except for Q_3 and Q_7, which have gate-source voltages of $2V_{eff}$ because they are biased at four times the current density. Thus, we find that the gate voltage of Q_3 equals

$$V_{G3} = 2V_{eff} + V_{tn} \tag{6.96}$$

and the drain-source voltage of Q_2 is given by

$$V_{DS2} = V_{S4} = V_{G3} - V_{GS4} = (2V_{eff} + V_{tn}) - (V_{eff} + V_{tn}) = V_{eff} \tag{6.97}$$

Therefore, Q_2 is biased on the edge of the triode region and the minimum output voltage is given by

$$V_{out} > V_{DS2} + V_{eff1} = 2V_{eff} \tag{6.98}$$

With the shown values for the W/L ratios, the power dissipation would be almost doubled over that of a classical cascode current mirror, assuming I_{bias} is set to one-seventh the nominal or maximum input current value, I_{in}. However, it is possible to bias the enhancement circuitry at lower densities and thereby save on power dissipation, albeit at the expense of speed, as the additional poles introduced by the enhancement circuitry would then be at lower frequencies.

A slightly modified version of the wide-swing enhanced output-impedance mirror is shown in Fig. 6.17. This variation obtains the bias voltage of the cascode transistor of the input diode-connected branch from the bias-generation circuitry of Fig. 6.3. This circuit suffers slightly with regard to large-signal dc matching, but has less power dissipation and area than the circuit of Fig. 6.16. Notice also that Q_2 of Fig. 6.16 has now been

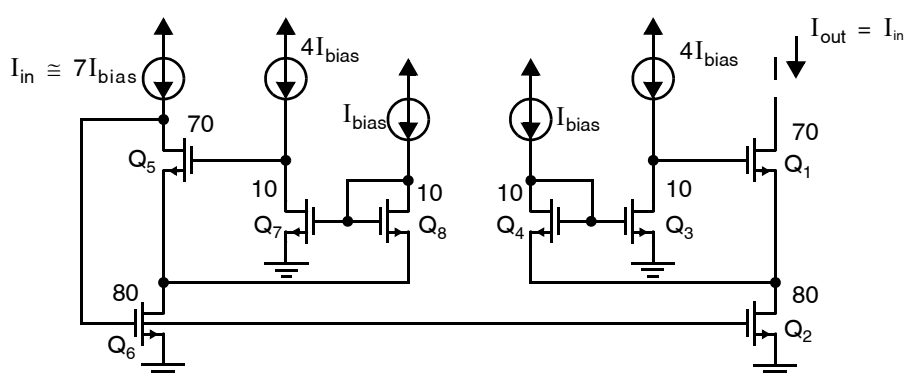


Fig. 6.16 A wide-swing current mirror with enhanced output impedance. Current source $7I_{bias}$ is typically set to the nominal or maximum input current, I_{in}.

separated into two transistors, Q_2 and Q_5, in Fig. 6.17. This change allows one to design an opamp without the output-impedance enhancement, but with wide-swing mirrors. Then, when output-impedance enhancement is desired, one simply includes the appropriate enhancement amplifiers and the only change required to the original amplifier is to connect the gates of the output cascode mirrors of the appropriate current mirrors to the enhancement amplifiers rather than the bias-generation circuit. Finally, it has been found through simulation that the modified circuit of Fig. 6.17 is less prone to instability than the circuit of Fig. 6.16.

It is predicted that this current mirror will become more important as the newer technologies force designers to use smaller power-supply voltages, which limit voltage swings. Also, the short-channel effects of newer technologies make the use of current mirrors with enhanced output impedance more desirable.

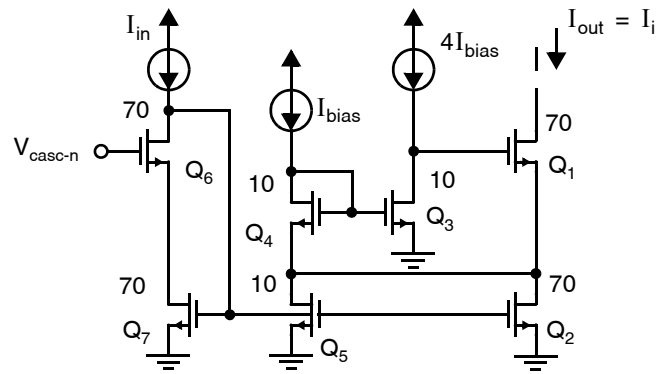


Fig. 6.17 A modified wide-swing current mirror with enhanced output impedance.

There are a couple of facts that designers should be aware of when they use output-impedance enhancement. The first fact is that sometimes it may be necessary to add local compensation capacitors to the enhancement loops to prevent ringing during transients. The second fact is that the inclusion of output-impedance enhancement can substantially slow down the settling times for large-signal transients. This is because during large-signal transients, the outputs of the enhancement loops can slew to voltages very different from those required after settling, and it can take a while for the outputs to return to the necessary voltages. A typical settling-time difference might be around a 50-percent increase. Whether this is worth the substantial increase in gain (perhaps as much as 30 dB or so) depends on the individual application.

6.3.4 Current-Mirror Symbol

It should now be clear that there are a number of different current-mirror circuits, each having various advantages and disadvantages, that can be used in a particular circuit application. Thus, from an architectural perspective, it is often desirable to describe a circuit without showing which particular current mirror is used. In these cases, we make use of the current-mirror symbol shown in Fig. 6.18(*a*). The arrow is on the input side of the mirror, which is the low-impedance side (with an impedance typically equal to $1/g_m$). The arrow also designates the direction of current flow on the input side. The ratio, 1:K, represents the current gain of the mirror. For example, the current mirror shown in Fig. 6.18(*a*) might be realized by the simple current mirror shown in Fig. 6.18(*b*). Finally, it should be mentioned that for illustrative purposes, some circuits in this book will be shown with a particular current mirror (such as the wide-swing mirror) to realize a circuit architecture. However, it should be kept in mind that similar circuits using almost any of the current mirrors just described (or described elsewhere) are possible.

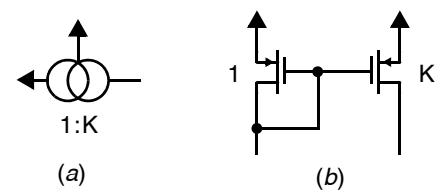


Fig. 6.18 (*a*) A symbol representing a current mirror. (*b*) Example of a simple current mirror.

6.4 FOLDED-CASCODE OPAMP

Many modern integrated CMOS opamps are designed to drive only capacitive loads. With such capacitive-only loads, it is not necessary to use a voltage buffer to obtain a low output impedance for the opamp. As a result, it is possible to realize opamps having higher speeds and larger signal swings than those that must also drive resistive loads. These improvements are obtained by having only a single high-impedance node at the output of an opamp that drives only capacitive loads. The admittance seen at all other nodes in these opamps is on the order of a transistor's transconductance, and thus they have relatively low impedance. By having all internal nodes of relatively low impedance, the speed of the opamp is maximized. It should also be mentioned that these low node impedances result in reduced voltage signals at all nodes other than the output node; however, the current signals in the various transistors can be quite large.[12] With these opamps, we shall see that the compensation is usually achieved by the load capacitance. Thus, as the load capacitance gets larger, the opamp usually becomes more stable but also slower. One of the most important parameters of these modern opamps is their transconductance value (i.e., the ratio of the output current to the input voltage). Therefore, some designers refer to these modern opamps as transconductance opamps, or *Operational Transconductance Amplifiers* (OTAs). A simple first-order small-signal model for an OTA is shown in Fig. 6.19. The amplifier's transconductance is G_{ma} and its frequency response in Fig. 6.19 is given simply by $G_{ma}Z_L$.

Key Point: *Operational transconductance amplifiers provide a large dc gain from a single stage. Their high output impedance makes them suitable for driving the mostly-capacitive loads encountered in CMOS integrated circuits. With feedback, they can also drive some resistive load.*

It should also be noted that feedback generally reduces the output impedance of a opamp circuit by a large factor. Hence, although the OTA itself has high output impedance, the output impedance of the overall feedback circuit is much lower, making it sometimes possible to drive resistive loads with OTAs, as long as the load resistance is modest.

An example of an opamp with a high-output impedance is the folded-cascode opamp, as shown in Fig. 6.20. The design shown is a differential-input single-ended output design. Note that all current mirrors in the circuit are wide-swing cascode current mirrors, discussed in Section 6.3.1. The use of these mirrors results in high-output impedance for the mirrors (compared to simple current mirrors), thereby maximizing the dc gain of the opamp. The basic idea of the folded-cascode opamp is to apply cascode transistors to the input differential pair but using transistors opposite in type from those used in the input stage. For example, the differential-pair transistors consisting of Q_1 and Q_2 are n-channel transistors in Fig. 6.20, whereas the cascode transistors consisting of Q_5 and Q_6 are p-channel transistors. This arrangement of opposite-type transistors allows the output of this single gain-stage amplifier to be taken at the same bias-voltage levels as the input signals. It should be mentioned that even though a folded-cascode amplifier is basically a single gain stage, its gain can be quite reasonable, on the order of 700 to 3,000. Such a high gain occurs because the gain is determined by the product of the input transconductance

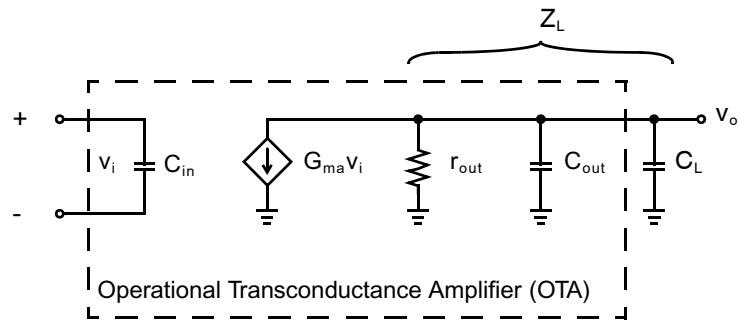


Fig. 6.19 First-order model of an operational transconductance amplifier (OTA) driving a capacitive load.

12. Because of their reduced voltage signals but large current signals, these types of opamps are sometimes referred to as *current-mode opamps*.

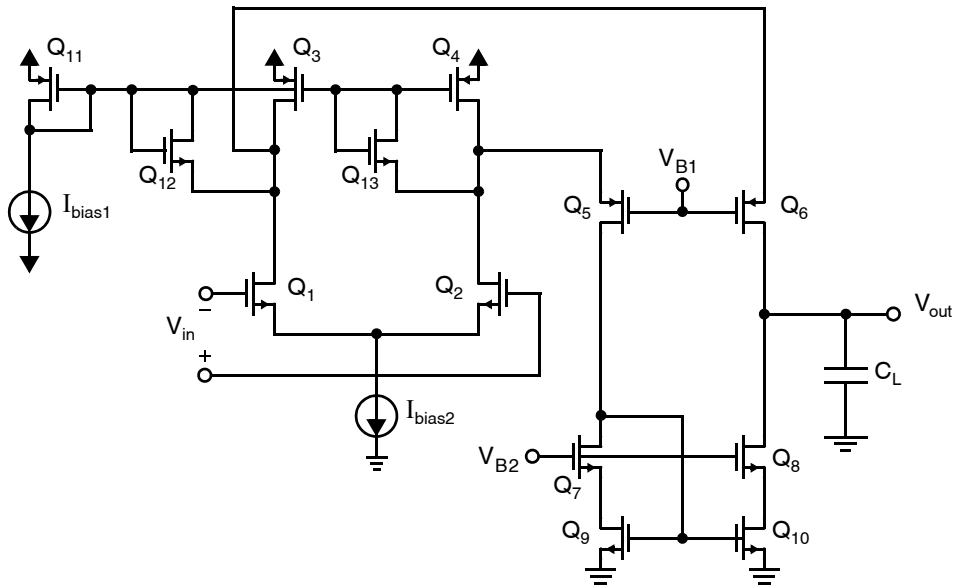


Fig. 6.20 A folded-cascode opamp.

and the output impedance, and the output impedance is quite high due to the use of cascode techniques. In technologies where transistor intrinsic gain is poor, auxiliary gain boosting amplifiers can be used to drive the gates of cascode devices $Q_{5\text{-}8}$ and increase dc gain as described in Section 6.3.2.

The shown differential-to-single-ended conversion is realized by the wide-swing current mirror composed of Q_7, Q_8, Q_9, and Q_{10}. In a differential-output design, these might be replaced by two wide-swing cascode current sinks, and common-mode feedback circuitry would be added, as is discussed in Sections 6.7 and 6.8.

It should be mentioned that a simplified bias network is shown so as not to complicate matters unnecessarily when the intent is to show the basic architecture. In practical realizations, Q_{11} and I_{bias1} might be replaced by the constant-transconductance bias network described in Chapter 7. In this case, V_{B1} and V_{B2} would be connected to $V_{casc\text{-}p}$ and $V_{casc\text{-}n}$, respectively, of Fig. 7.8 to maximize the output voltage signal swing.

An important addition of the folded-cascode opamp shown is the inclusion of two extra transistors, Q_{12} and Q_{13}. These two transistors serve two purposes. One is to increase the slew-rate performance of the opamp, as will be discussed in Example 6.9. However, more importantly, during times of slew-rate limiting, these transistors prevent the drain voltages of Q_1 and Q_2 from having large transients where they change from their small-signal voltages to voltages very close to the negative power-supply voltage. Thus the inclusion of Q_{12} and Q_{13} allows the opamp to recover more quickly following a slew-rate condition.

The compensation is realized by the load capacitor, C_L, and realizes dominant-pole compensation. In applications where the load capacitance is very small, it is necessary to add additional compensation capacitance in parallel with the load to guarantee stability. If lead compensation is desired, a resistor can be placed in series with C_L. While lead compensation may not be possible in some applications, such as when the compensation capacitance is mostly supplied by the load capacitance, it is more often possible than many designers realize (i.e., it is often possible to include a resistor in series with the load capacitance).

Key Point: *The folded-cascode opamp is a transconductance amplifier whose frequency response is dominated by the output pole at* $1/R_{out}C_L$.

The bias currents for the input differential-pair transistors are equal to $I_{bias2}/2$. The bias current of one of the p-channel cascode transistors, Q_5 or Q_6, and hence the transistors in the output-summing current mirror as well, is

equal to the drain current of Q_3 or Q_4 minus $I_{bias2}/2$. This drain current is established by I_{bias1} and the ratio of $(W/L)_3$, or $(W/L)_4$, to $(W/L)_{11}$. Since the bias current of one of the cascode transistors is derived by a current subtraction, for it to be accurately established, it is necessary that both I_{bias1} and I_{bias2} be derived from a single-bias network. In addition, any current mirrors used in deriving these currents should be composed of transistors realized as parallel combinations of unit-size transistors. This approach eliminates inaccuracies due to second-order effects caused by transistors having non-equal widths.

6.4.1 Small-Signal Analysis

In a small-signal analysis of the folded-cascode amplifier, it is assumed that the differential output current from the drains of the differential pair, Q_1, Q_2, is applied to the load capacitance, C_L. Specifically, the small-signal current from Q_1 passes directly from the source to the drain of Q_6 and thus C_L, while the current from Q_2 goes indirectly through Q_5 and the current mirror consisting of Q_7 to Q_{10}.[13] Although these two paths have slightly different transfer functions due to the poles caused by the current mirror, when an n-channel transistor current mirror is used, a pole-zero doublet occurs at frequencies quite a bit greater than the opamp unity-gain frequency and can usually be ignored. Hence, ignoring the high-frequency poles and zero, an approximate small-signal transfer function for the folded-cascode opamp is given by

$$A_V = \frac{V_{out}(s)}{V_{in}(s)} = g_{m1}Z_L(s) \tag{6.99}$$

In this case, the amplifier's transconductance gain is simply the transconductance of each of the transistors in the input differential pair, $G_{ma} = g_{m1}$, and $Z_L(s)$ is the impedance to ground seen at the output node. This impedance consists of the parallel combination of the output load capacitance, the impedance of any additional network added for stability, and the output impedance of the opamp (i.e., looking into the drains of Q_6 and Q_8). Thus, neglecting other parasitic poles, we have

$$A_V = \frac{g_{m1}r_{out}}{1 + sr_{out}C_L} \tag{6.100}$$

where r_{out} is the output impedance of the opamp. This impedance is quite high, on the order of $g_m r_{ds}^2/2$ or greater if output-impedance enhancement is used.

For mid-band and high frequencies, the load capacitance dominates, and we can ignore the unity term in the denominator and thus have

$$A_V \cong \frac{g_{m1}}{sC_L} \tag{6.101}$$

from which the unity-gain frequency of the opamp is found to be

$$\omega_{ta} = \frac{g_{m1}}{C_L} \tag{6.102}$$

Key Point: *If the bandwidth of a folded-cascode opamp is insufficient, one must increase the input transconductance.*

With feedback, the loop unity-gain frequency is $\omega_t = \beta g_{m1}/C_L$. Therefore, for large load capacitances, the unity-gain frequency is much less than the second- and higher-order poles, and the phase margin, even with $\beta = 1$, is close to 90°. Hence, amplifier bandwidth is maximized by maximizing the transconductance of the input transistors, which in turn is maximized by using wide n-channel devices and ensuring that the input transistor pair's bias current is substantially larger than the bias current of the cascode

13. This is based on the assumption that g_{m5} and g_{m6} are much larger than g_{ds3} and g_{ds4}.

transistors and current mirror. Note that this approach also maximizes the dc gain (i.e., $g_{m1}r_{out}$) since not only does it maximize g_{m1}, but it also maximizes r_{out} since all transistors connected to the output node are biased at lower current levels (for a given total power dissipation). A practical upper limit on the ratio of the bias currents of the input transistors to the currents of the cascode transistors might be around four or so. If too high a ratio is used, the bias currents of the cascode transistors cannot be well established since they are derived by current subtractions. Another advantage of having very large transconductances for the input devices is that the thermal noise due to this input pair is reduced. Thus, since much of the bias current in folded-cascode opamps flows through the input differential pair, these opamps often have a better thermal noise performance than other opamp designs having the same power dissipation.

If the unity-gain frequency (6.102) is high enough, the second- and higher-order poles may become significant. The second poles of this opamp are primarily due to the time constants introduced by the impedance and parasitic capacitances at the sources of the p-channel cascode transistors, Q_5 and Q_6. The impedances at these nodes is one over the transconductance of the cascode transistors.[14] Since p-channel transistors are used here, possibly biased at lower currents, these impedance are typically substantially greater than the source impedance of most n-channel transistors in the signal path. As a consequence, when high-frequency operation is important, these impedances can be reduced by making the currents in the p-channel cascode transistors around the same level as the bias currents of the input transistors. The parasitic capacitance at the sources of the cascode transistors is primarily due to the gate-source capacitances of the cascode transistors as well as the drain-to-bulk and drain-to-gate capacitances of the input transistors and the current-source transistors Q_3 and Q_4. Therefore, minimizing junction areas and peripheries at these two nodes is important.

In summary, if the phase margin of a folded-cascode opamp with feedback is insufficient, one has two choices:

a. Add an additional capacitance in parallel with the load to decrease the dominant pole. This improves phase margin without additional power consumption and without effecting the dc gain, but decreases amplifier bandwidth and increases area.

b. Increase the current and device widths in the output stage, I_{D5} and I_{D6}. This will increase the second pole frequency and improve phase margin, but sacrifices dc gain and increases power consumption.

Key Point: *If the phase margin of a folded-cascode opamp is insufficient, one must either reduce the bandwidth by adding output capacitance, or burn more power in the output transistors to increase the second pole frequency.*

Note that we have already seen in Chapter 4 that an upper limit on the second pole frequency of a cascode gain stage is imposed by the intrinsic speed (unity-gain frequency) of the cascode transistor—in this case $Q_{5,6}$.

$$f_{p2} < \frac{3\mu_p V_{eff5}}{4\pi L_5^2} = f_{T5} \tag{6.103}$$

Hence, option (b) above is ultimately subject to this constraint.

It is also possible to employ lead compensation. When the load is a series combination of a compensation capacitance, C_L, and a lead resistance, R_C, the small-signal transfer function is given by

$$A_V = \frac{g_{m1}}{\dfrac{1}{r_{out}} + \dfrac{1}{R_C + 1/sC_L}} \cong \frac{g_{m1}(1 + sR_CC_L)}{sC_L} \tag{6.104}$$

where the approximate term is valid at mid and high frequencies. Here we see that, as in Section 6.2, R_C can be chosen to place a zero at 1.7 times the unity-gain frequency.

---

14. This is true only at high frequencies, where the output impedance of the amplifier is small due to the load and/or compensation capacitance. This is the case of interest here. At low frequencies, the impedance looking into the source of Q_6 is given by $R_{s6} = 1/g_{m6} + R_L/(g_{m6}r_{ds6})$, where R_L is the resistance seen looking out the drain of Q_6.

6.4.2 Slew Rate

The diode-connected transistors, Q_{12} and Q_{13}, are turned off during normal operation and have almost no effect on the opamp. However, they substantially improve the operation during times of slew-rate limiting [Law, 1983]. To appreciate their benefit, consider first what happens during times of slew-rate limiting when they are not present. Assume there is a large differential input voltage that causes Q_1 to be turned on hard and Q_2 to be turned off. Since Q_2 is off, all of the bias current of Q_4 will be directed through the cascode transistor Q_5, through the n-channel current mirror, and out of the load capacitance. Thus, the output voltage will decrease linearly with a slew-rate given by

$$SR = \frac{I_{D4}}{C_L} \tag{6.105}$$

Also, since all of I_{bias2} is being diverted through Q_1, and since this current is usually designed to be greater than I_{D3}, both Q_1 and the current source I_{bias2} will go into the triode region, causing I_{bias2} to decrease until it is equal to I_{D3}. As a result, the drain voltage of Q_1 approaches that of the negative power-supply voltage. When the opamp is coming out of slew-rate limiting, the drain voltage of Q_1 must slew back to a voltage close to the positive power supply before the opamp operates in its linear region again. This additional slewing time greatly increases the distortion and also increases the transient times during slew-rate limiting (which occurs often for opamps used in switched-capacitor applications).

Next, consider the case where the diode-connected transistors, Q_{12} and Q_{13}, are included. *Their main purpose is to clamp the drain voltages of* Q_1 *or* Q_2 *so they don't change as much during slew-rate limiting*. A second, more subtle effect dynamically increases the bias currents of both Q_3 and Q_4 during times of slew-rate limiting. This increased bias current results in a larger maximum current available for charging or discharging the load capacitance. To see this increase in bias current, consider a case similar to the one just described, where a large differential input causes Q_1 to be fully on while Q_2 is off. In this case, the diode-connected transistor Q_{12} conducts with the current through Q_{12} coming from the diode-connected transistor Q_{11}. Thus, the current in Q_{11} increases, causing the currents in bias transistors Q_3 and Q_4 to also increase, until the sum of the currents of Q_{12} and Q_3 are equal to the bias current I_{bias2}. Note that the current in Q_4 also increases since it is equal to the current in Q_3. This increase in bias current of Q_4 results in an increase of the maximum current available for discharging C_L. In summary, not only are the voltage excursions less, but the maximum available current for charging or discharging the load capacitance is also greater during times of slew-rate limiting.

EXAMPLE 6.9

Find reasonable transistor sizes for the folded-cascode opamp shown in Fig. 6.20 to satisfy the following design parameters. Also find the opamp's unity-gain frequency (without feedback) and slew rate, both without and with the clamp transistors.

- Assume the process parameters for the 0.18-μm process in Table 1.5, a single 1.8-V power supply, and limit the current dissipation of the opamp to 0.4 mA.
- Set the ratio of the current in the input transistors to that of the cascode transistors to be 4:1. Also, set the bias current of Q_{11} to be 1/10th that of Q_3 (or Q_4) such that its current can be ignored in the power dissipation calculation.
- The maximum transistor width should be 180 μm and channel lengths of 0.4 μm should be used in all transistors.
- All transistors should have effective gate-source voltages of around 0.24 V except for the input transistors, whose widths should be set to the maximum value of 180 μm. Also, round all transistor widths to the closest multiple of 2 μm, keeping in mind that if a larger transistor is to be matched to a smaller one, the larger transistor should be built as a parallel combination of smaller transistors.
- Finally, assume the load capacitance is given by $C_L = 2.5$ pF.

Solution

The total current in the opamp, I_{total}, excluding the current in the bias network, is equal to $I_{D3} + I_{D4}$, which is equal to $2(I_{D1} + I_{D6})$. Defining $I_B \equiv I_{D5} = I_{D6}$ and noting that we are asked to make $I_{D1} = 4I_{D6}$, we have

$$I_{total} = 2(I_{D1} + I_{D6}) = 2(4I_B + I_B) = 10I_B \tag{6.106}$$

Since the current dissipation is specified to be 0.4 mA, we have

$$I_B = I_{D5} = I_{D6} = \frac{I_{total}}{10} = 40\ \mu A \tag{6.107}$$

This bias current for I_B implies that $I_{D3} = I_{D4} = 5I_{D5} = 200\ \mu A$, and $I_{D1} = I_{D2} = 4I_{D5} = 160\ \mu A$. Now, we let all transistor channel lengths be $0.4\ \mu m$, about $2\times$ larger than the minimum gate length in this technology. This choice allows us to immediately determine the sizes of most transistors using

$$\left(\frac{W}{L}\right)_i = \frac{2I_{Di}}{\mu_i C_{ox} V_{effi}^2} \tag{6.108}$$

SPICE! *Simulate this opamp with the provided netlist.*

and then round to the closest multiple of $2\ \mu m$ in transistor widths. Input transistors Q_1 and Q_2 were an exception; their widths are set at the prescribed maximum of $180\ \mu m$. This puts those devices on the border of subthreshold operation, thereby maximizing their transconductance for the given bias current. Table 6.1 lists reasonable values for the resulting dimensions of all transistors. Note that the larger widths are exactly divisible by the smaller widths of a transistor of the same type and thus allows larger transistors to be realized as a parallel combination of smaller transistors. The width of Q_{11} was determined from the requirement that $I_{D11} = I_{D3}/10 = 20\ \mu A = I_{bias,1}$. The widths of Q_{12} and Q_{13} were somewhat arbitrarily chosen to equal the width of Q_{11}.

Under a square-law assumption, the transconductance of the input transistors would be given by

$$\sqrt{2I_{D1}\mu_n C_{ox}(W/L)_1} = 6.24\ \text{mA/V} \tag{6.109}$$

However, this is greater than the maximum transconductance achieved in subthreshold operation.

$$g_{m1(sub-th)} = \frac{qI_{D1}}{nkT} = 4\ \text{mA/V} \tag{6.110}$$

The maximum value computed in (6.110) is assumed as an approximation, although the real transconductance will be less than this upper limit. The unity-gain frequency of the opamp is given by

$$\omega_{ta} = \frac{g_{m1}}{C_L} = 1.6\times10^9\ \text{rad/s} \Rightarrow f_{ta} = 255\ \text{MHz} \tag{6.111}$$

Table 6.1 **The transistor dimensions (in μm) of the opamp of Fig. 6.20.**

Q_1	180/0.4	Q_6	8/0.4	Q_{11}	4/0.4
Q_2	180/0.4	Q_7	2/0.4	Q_{12}	4/0.4
Q_3	40/0.4	Q_8	2/0.4	Q_{13}	4/0.4
Q_4	40/0.4	Q_9	2/0.4		
Q_5	8/0.4	Q_{10}	2/0.4		

The slew rate without the clamp transistors is given by

$$SR = \frac{I_{D4}}{C_L} = 80\ V/\mu s \tag{6.112}$$

When the clamp transistors are included, during slew-rate limiting, we have

$$I_{D12} + I_{D3} = I_{bias2} \tag{6.113}$$

But

$$I_{D3} = 10 I_{D11} \tag{6.114}$$

and

$$I_{D11} = 20\ \mu A + I_{D12} \tag{6.115}$$

Substituting (6.114) and (6.115) into (6.113) and solving for I_{D11} gives

$$I_{D11} = \frac{I_{bias2} + 20\ \mu A}{11} \tag{6.116}$$

which implies that the value of I_{D11} during slew-rate limiting is 30.9 μA and $I_{D3} = I_{D4} = 10 I_{D11} = 0.309$ mA, which is substantially larger than the slew current available without the clamp transistors. This larger bias current will give a slew-rate value of

$$SR = \frac{I_{D4}}{C_L} = 124\ V/\mu s \tag{6.117}$$

More importantly, the time it takes to recover from slew-rate limiting will be substantially decreased.

EXAMPLE 6.10

The opamp in Example 6.9 is simulated and found to have an equivalent second pole frequency at $\omega_{eq} = 2\pi \cdot 365$ MHz. Select a value for the lead compensation resistor, R_C, to provide a phase margin of 85° when the opamp is used in a unity-gain configuration.

Solution

The phase margin without R_C is given by

$$\text{Phase Margin} = 90° - \tan^{-1}\left(\frac{\omega_t}{\omega_{eq}}\right) = 90° - \tan^{-1}\left(\frac{2\pi \cdot 255\ \text{MHz}}{2\pi \cdot 365\ \text{MHz}}\right) = 55°$$

In order to increase this by $30° = \tan^{-1}(1/1.7)$, the lead compensation zero must be placed at $1/R_C C_L = 1.7\omega_t$. Since $\beta = 1$, we have $\omega_t = \omega_{ta}$. Hence, a reasonable size for R_C in series with C_L is given by

$$R_C = \frac{1}{1.7 C_L \omega_t} = \frac{1}{1.7 g_{m1}} = 147\ \Omega \tag{6.118}$$

6.5 CURRENT MIRROR OPAMP

Another popular opamp often used when driving on-chip capacitive-only loads is the current mirror opamp shown in simplified form in Fig. 6.21. It is immediately seen that all nodes are low impedance except for the output node. By using good current mirrors having high output impedance, a reasonable overall gain can be achieved. A more detailed example of a current mirror opamp is shown in Fig. 6.22.

The overall transfer function of this opamp will closely approximate dominant-pole operation. In a similar analysis to that given for the folded-cascode opamp, we have

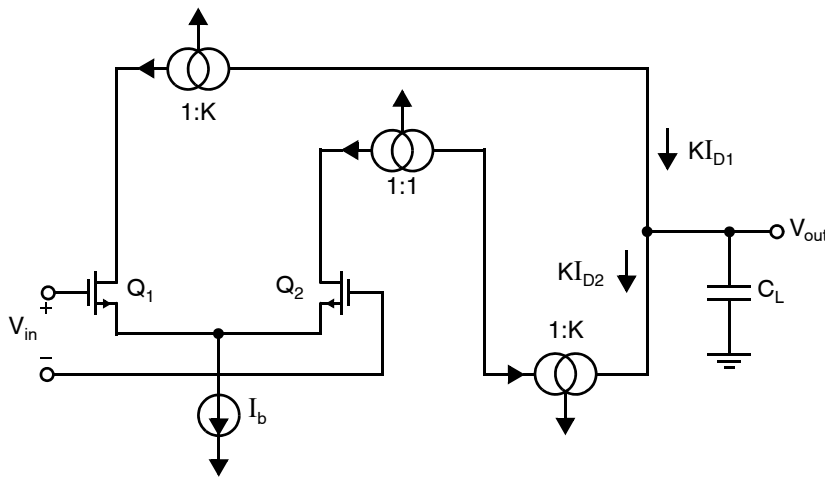


Fig. 6.21 A simplified current-mirror opamp.

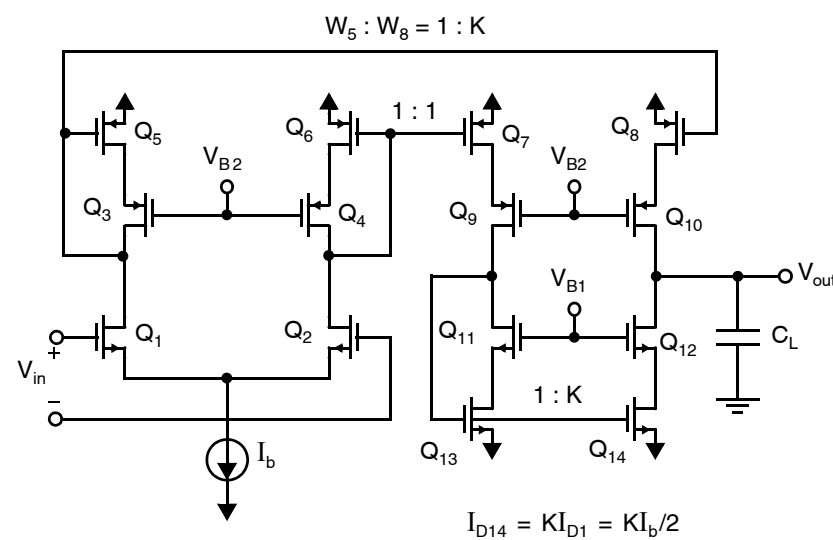


Fig. 6.22 A current-mirror opamp with wide-swing cascode current mirrors.

$$A_V = \frac{V_{out}(s)}{V_{in}(s)} = Kg_{m1}Z_L(s) = \frac{Kg_{m1}r_{out}}{1 + sr_{out}C_L} \cong \frac{Kg_{m1}}{sC_L} \tag{6.119}$$

The K factor is the current gain from the input transistors to the output sides of the current mirrors connected to the output node. Hence, the amplifier's overall transconductance is $G_{ma} = Kg_{m1}$. Using (6.119), we can solve for the opamp's unity-gain frequency (without feedback), resulting in

$$\omega_{ta} = \frac{Kg_{m1}}{C_L} = \frac{2KI_{D1}}{C_L V_{eff,1}} \tag{6.120}$$

If the power dissipation is specified, the total current,

$$I_{total} = (3 + K)I_{D1} \tag{6.121}$$

is known for a given power-supply voltage. Substituting (6.121) into (6.120), we obtain

$$\omega_{ta} = \frac{K}{(3 + K)} \frac{2I_{total}}{V_{eff,1}C_L} \tag{6.122}$$

Equations (6.120) and (6.122) assume a square law for input transistors $Q_{1,2}$ and emphasize the desire to bias the input pair with low values of V_{eff}. If they are biased with very low V_{eff}, their transconductance approaches the upper limit obtained in subthreshold operation, $g_{m1(sub-th)} \cong qI_{D1}/nkT$ in which case

$$\omega_{ta} = \frac{Kq\left(\frac{I_{total}}{3 + K}\right)}{nkTC_L} = \frac{K}{(3 + K)} \frac{qI_{total}}{nkTC_L} \tag{6.123}$$

Regardless, the opamp's transconductance (i.e., Kg_{m1}) increases with increasing K, and therefore the unity-gain frequency is also larger. This simple result assumes the unity-gain frequency is limited by the load capacitance rather than any high-frequency poles caused by the time constants of the internal nodes. A practical upper limit on K might be around five. The use of large K values also maximizes the gain for I_{total} fixed since r_{out} is roughly independent of K for large K.

In the circuit shown in Fig. 6.22, the important nodes for determining the non-dominant poles are the drain of Q_1, primarily, and the drains of Q_2 and Q_9, secondly. Increasing K increases the capacitances of these nodes while also increasing the equivalent resistances.[15] As a result, the equivalent second pole moves to lower frequencies. If K is increased too much, the second pole frequency will approach ω_{ta}, and an increase in C_L will be required to decrease ω_{ta} and maintain stability, unless β is very small. Thus, increasing K decreases the bandwidth when the equivalent second pole dominates. This is most likely to occur when the load capacitance is small. If it is very important that speed is maximized, K might be taken as small as one. A reasonable compromise for a general-purpose opamp might be $K = 2$.

The slew rate of the current-mirror opamp is found by assuming there is a very large input voltage. In this case, all of the bias current of the first stage will be diverted through either Q_1 or Q_2. This current will be amplified by the current gain from the input stage to the output stage, (i.e., K). Thus, during slew-rate limiting, the total current available to charge or discharge C_L is KI_b. The slew rate is therefore given by

$$SR = \frac{KI_b}{C_L} \tag{6.124}$$

15. Increasing K implies that the currents at the input sides of the current mirrors are smaller, for a given total power dissipation. Also, the widths of transistors at the input sides of current mirrors will be smaller. Both of these effects cause the transconductances of transistors at the input side of current mirrors to be smaller, and therefore the impedances (in this case, $1/g_m$) will be larger.

For a given total power dissipation, this slew rate is maximized by choosing a large K value. For example, with $K = 4$ and during slew-rate limiting, 4/5 of the total bias current of the opamp will be available for charging or discharging C_L. This result gives a current-mirror opamp superior slew rates when compared to a folded-cascode opamp, even when the clamp transistors have been included in the folded-cascode opamp. Also, there are no problems with large voltage transients during slew-rate limiting for the current-mirror opamp.

Key Point: *For a current-mirror opamp, when the load capacitance is large and the output pole clearly dominant, a large current mirror ratio, K, maximizes gain and bandwidth. If the second pole frequency becomes significant, lower values of K are required to keep the second pole frequency above the unity-gain frequency and maintain a healthy phase margin.*

In summary, due primarily to the larger bandwidth and slew rate, the current-mirror opamp is often preferred over a folded-cascode opamp. However, it will suffer from larger thermal noise when compared to a folded-cascode amplifier because its input transistors are biased at a lower proportion of the total bias current and therefore have a smaller transconductance. As seen in the chapter on noise, a smaller transconductance results in larger thermal noise when the noise is referred back to the input of the opamp.

EXAMPLE 6.11

SPICE! *Simulate this opamp with the provided netlist.*

Assume the current-mirror opamp shown in Fig. 6.22 has all transistor lengths equal to 0.4 μm and transistor widths as given in Table 6.2. Notice that $K = 2$. Assume the process parameters for the 0.18-μm process in Table 1.5, a single 1.8-V power supply, and that the opamp's total current dissipation is 0.4 mA. The load capacitance is $C_L = 2.5$ pF.

Find the slew rate and the unity-gain frequency, assuming the equivalent second pole does not dominate. Estimate the equivalent second pole. Would it be necessary to increase C_L if a 75° phase margin were required with $\beta = 1$ and without using lead compensation? What if lead compensation were used?

Solution

Since the total bias current is given by $(3 + K)I_b/2$, we have

$$I_b = \frac{2I_{total}}{(3+K)} = \frac{2(0.4\ \text{mA})}{(3+K)} = 160\ \mu\text{A} \tag{6.125}$$

Thus, the bias currents of all transistors, except those in the output stage, are 80 μA. The bias currents of transistors in the output stage are twice that of the input stage, which is 160 μA. The effective gate-source voltage of the input transistors may be shown to be less than 100 mV, so their transconductance is near the limit obtained in subthreshold.

$$g_{m1(sub-th)} = \frac{qI_{D1}}{nkT} = 2\ \text{mA/V} \tag{6.126}$$

Table 6.2 The transistor sizes (in μm) for the opamp used in Example 6.11.

Q_1	90/0.4	Q_7	20/0.4	Q_{13}	4/0.4
Q_2	90/0.4	Q_8	40/0.4	Q_{14}	8/0.4
Q_3	20/0.4	Q_9	20/0.4		
Q_4	20/0.4	Q_{10}	40/0.4		
Q_5	20/0.4	Q_{11}	4/0.4		
Q_6	20/0.4	Q_{12}	8/0.4		

The transconductance of the opamp will be K times this or 4 mA/V . The opamp unity-gain frequency is now given by

$$\omega_{ta} = \frac{Kg_{m1}}{C_L} = 1.6\times10^9 \text{ rad/s} \Rightarrow f_{ta} = 255 \text{ MHz} \tag{6.127}$$

This is the same result obtained for the folded-cascode opamp of the previous example with the same current consumption and load capacitance. If a higher value of K were chosen, the current-mirror opamp would provide superior bandwidth.

Continuing, the slew rate is given by

$$SR = \frac{KI_b}{C_L} = 128 \text{ V}/\mu\text{s} \tag{6.128}$$

which compares favorably to 80 V/μs for the folded-cascode opamp without the clamp transistors and slightly higher than the folded-cascode opamp with the clamp transistors.

When estimating the equivalent second pole, the junction capacitances will be ignored to simplify matters. In reality, the drain-to-bulk capacitances of the input transistors could contribute a significant factor, and this amount can be determined using simulation. The dominant node almost certainly will occur at the drain of Q_1. The impedance at this node is given by

$$R_1 = 1/g_{m5} = 1.3 \text{ k}\Omega \tag{6.129}$$

In addition, the capacitance will be primarily due to the gate-source capacitances of Q_5 and Q_8. We therefore have

$$C_1 = C_{gs5} + C_{gs8} = (1 + K)C_{gs5} \tag{6.130}$$

resulting in

$$C_1 = (1 + K)(2/3)C_{ox}W_5 L_5 = 0.14 \text{ pF} \tag{6.131}$$

With these values of impedance, the time constant for this node is given by

$$\tau_1 = R_1C_1 = 0.18 \text{ ns} \tag{6.132}$$

In a similar manner, we can calculate the impedances and, hence, the time constant for the drain of Q_2 to be $R_2 = 1.3$ kΩ, $C_2 = 0.091$ pF, and $\tau_2 = R_2C_2 = 0.12$ ns. The other important time constant comes from the parasitic capacitors at the drain of Q_9 . Here, we have

$$R_3 = 1/g_{m13} = 1.5 \text{ k}\Omega \tag{6.133}$$

and

$$C_3 = C_{gs13} + C_{gs14} = (1 + K)(2/3)C_{ox}W_{13} L_{13} = 0.027 \text{ pF} \tag{6.134}$$

resulting in $\tau_3 = 0.04$ ns. The time constant of the equivalent second pole can now be estimated to be given by

$$\tau_{2eq} = \tau_1 + \tau_2 + \tau_3 = 0.34 \text{ ns} \tag{6.135}$$

This result gives an equivalent second pole at the approximate frequency of

$$p_{2eq} = \frac{1}{\tau_{2eq}} = 2.94\times10^9 \text{ rad/s} = 2\pi \times 468 \text{ MHz} \tag{6.136}$$

If lead compensation is not used and a 75° phase margin is required, the unity-gain frequency must be constrained to be less than 0.27 times the equivalent second pole from Table 5.1 of Chapter 5. Specifically, the unity-gain

frequency must be less than 126 MHz . To achieve this unity-gain frequency, we need to increase C_L by the factor 255 MHz/(126 MHz), resulting in C_L being 5 pF.

If lead compensation is to be used, the unity-gain frequency can be chosen so that only a 55° phase margin is achieved *before* the lead resistor is added. This approach would allow the unity-gain frequency to be as high as 0.7 times the equivalent second-pole frequency, which is greater than the value computed in (6.127). The benefits of lead compensation are obvious: the unity-gain frequency is increased by 150% and no compensation capacitor is required, reducing the physical area of the circuit.

As a final comment, notice that the delay through the signal path from the drain of Q_2 to the output (i.e., $\tau_2 + \tau_3 = 0.16$ ns) is approximately the same as the delay through the signal path from the drain of Q_1 to the output, which is $\tau_1 = 0.18$ ns . This near equivalence helps to minimize harmful effects caused by the pole-zero doublet introduced at high frequencies (i.e. around the frequency of the second pole) by the existence of two signal paths.

6.6 LINEAR SETTLING TIME REVISITED

We saw in Chapter 5 that the time constant for linear settling time was equal to the inverse of $\omega_{-3\ dB}$ for the closed-loop circuit gain. We also saw that $\omega_{-3\ dB}$ is given by the relationship

$$\omega_{-3\ dB} = \omega_t \tag{6.137}$$

However, while for the classical two-stage CMOS opamp the unity-gain frequency remains relatively constant for varying load capacitances, the unity-gain frequencies of the folded-cascode and current-mirror amplifiers are strongly related to their load capacitance. As a result, their settling-time performance is affected by both the feedback factor as well as the effective load capacitance.

For the folded-cascode opamp, its unity-gain frequency is given by

$$\omega_{ta} = \frac{g_{m1}}{C_L} \tag{6.138}$$

whereas for a current-mirror opamp, it is equal to

$$\omega_{ta} = \frac{Kg_{m1}}{C_L} \tag{6.139}$$

Thus we see that for both these high-output impedance opamps, their unity-gain frequency is inversely proportional to the load capacitance. To determine the –3-dB frequency of a closed-loop opamp, consider the general case shown in Fig. 6.23. At the opamp output, C_{load} represents the capacitance of the next stage that the opamp must drive, while C_C is a compensation capacitance that might be added to maintain a sufficient phase margin. At the input side, C_p represents parasitic capacitance due to large transistors at the opamp input as well as any switch capacitance.

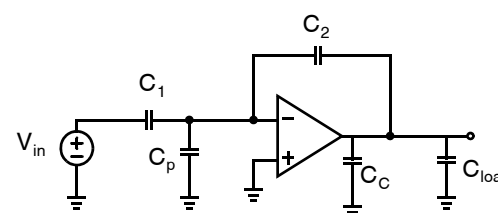


Fig. 6.23 A general circuit for determining the –3-dB frequency of a closed-loop amplifier.

Defining $\omega_t = \beta\omega_{ta}$, the feedback network is due to capacitances C_1, C_2 , and C_p resulting in

$$\beta = \frac{1/[s(C_1 + C_p)]}{1/[s(C_1 + C_p)] + 1/(sC_2)} = \frac{C_2}{C_1 + C_p + C_2} \tag{6.140}$$

To determine the unity-gain frequency it is necessary to consider the circuit's open-loop operation. Specifically, treating the negative opamp input as an open circuit, the effective load capacitance, C_L, is seen to be given by the parallel combination of C_C and C_{load}, as well as that seen looking into C_2. The capacitance seen looking into C_2 is equal to the series combination of C_2 together with $C_1 + C_p$. Combining all these capacitances together, we have

$$C_L = C_C + C_{load} + \frac{C_2(C_1 + C_p)}{C_1 + C_p + C_2} \tag{6.141}$$

EXAMPLE 6.12

Consider the current-mirror opamp with no lead compensation in Example 6.11 being used in the circuit shown in Fig. 6.23 with $C_1 = C_2 = C_C = C_{load} = 5$ pF. What is the linear settling time required for 0.1 percent accuracy with?

Solution

First, the gate-source capacitances of the input devices of the current-mirror opamp can be calculated to be

$$C_{gs1} = (2/3) \times 90 \times 0.4 \times 8.5 \text{ fF}/\mu\text{m}^2 = 0.21 \text{ pF} \tag{6.142}$$

The capacitance seen looking into the inverting input of the opamp is one-half this value since the gate-source capacitances of the two input devices are in series. Thus, the parasitic capacitance, C_P, is 0.11 pF. Therefore, the effective load capacitance is given by

$$C_L = 5 + 5 + \frac{5(5 + 0.11)}{5 + 5 + 0.11} = 12.53 \text{ pF} \tag{6.143}$$

which results in

$$\omega_{ta} = \frac{Kg_{m1}}{C_L} = \frac{2 \times 2 \text{ mA/V}}{12.53 \text{ pF}} = 3.19 \times 10^8 \text{ rad/s} \tag{6.144}$$

or equivalently,

$$f_{ta} = 50.8 \text{ MHz} \tag{6.145}$$

Now, the feedback factor, β, is seen to be

$$\beta = \frac{5}{5 + 0.11 + 5} = 0.49 \tag{6.146}$$

resulting in

$$\tau = \frac{1}{\beta\omega_{ta}} = 6.4 \text{ ns} \tag{6.147}$$

Finally, for 0.1 percent accuracy, we need a linear settling time of 7τ or 45 ns.

6.7 FULLY DIFFERENTIAL OPAMPS

Key Point: *Fully differential amplifiers generally have similar power consumption, but twice the available signal swing of their single-ended counterparts. They can also exhibit similar small signal gain and bandwidth.*

There is often little difference between the circuits required to process single-ended and fully differential signals. In particular, Fig. 6.24 compares fully differential pair circuit its single-ended output counterpart. The only difference them is that a current mirror load is replaced by two matched current sources in the fully differential circuit. Notice that the power dissipated in these two circuits is the same. Since the voltage swing on any individual node in a circuit is usually limited by fixed supply and bias voltages, fully differential circuits have twice the available signal swing of their single-ended counterparts because they make full use of the swing available at two circuit nodes instead of just one. Also, note that the single-ended circuit has a slew rate equal to the total available bias current over the load capacitance, I/C_L, whereas in the fully differential circuit each side has $I/2$ available when slew rate limited and so the slew rate of the differential voltage v_o is also $I/2C_L - (-I/2C_L) = I/C_L$. Although Fig. 6.24 illustrates these points for an actively-loaded differential pair, they are also true for more complex amplifiers, such as the folded cascode opamp.

The small signal performance of a fully differential amplifier is in many way equivalent to that of a single-ended amplifier with similar power consumption. Fig. 6.25 illustrates the point for single-stage amplifiers. Assuming that the circuit's input stage is in both cases a differential pair under the same dc bias, and each half of the fully differential output has an output and load impedance similar to that of the single-ended output, the circuits have similar small-signal gain and bandwidth.

One of the major driving forces behind the use of fully differential signals is to help reject common mode noise. Fig. 6.26 illustrates the impact of additive noise on fully differential signals. Common mode noise, n_{cm}, appears identically on both half-signals, and is therefore cancelled when only the difference between them is considered (for example, when the noisy signal v_o' serves as the differential input to a circuit with good common-mode rejection). In many analog integrated circuits, the largest noise sources are due to fluctuations in supply and bias voltages and from pass-transistor switches turning off in switched-capacitor applications. These appear as common mode noise and are

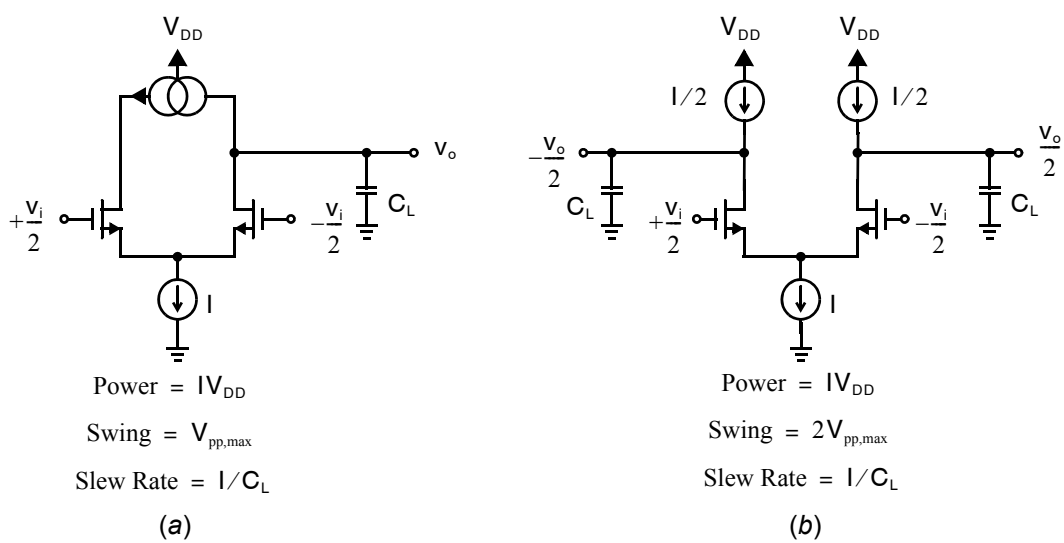
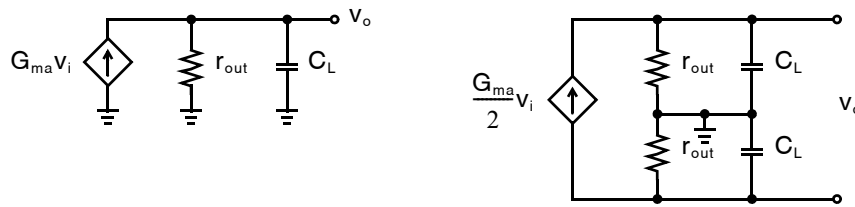


Fig. 6.24 A comparison of simple actively loaded differential pairs: (*a*) single-ended output; (*b*) fully differential.



dc gain: $G_{ma}r_{out}$

$\omega_{-3dB} = 1/(r_{out}C_L)$

$\omega_{ta} = G_{ma}/C_L$

(*a*)

dc gain: $\frac{G_{ma}}{2}2r_{out} = G_{ma}r_{out}$

$\omega_{-3dB} = 1/(r_{out}C_L)$

$\omega_{ta} = G_{ma}/C_L$

(*b*)

Fig. 6.25 Equivalent (*a*) single-ended and (*b*) fully differential small-signal amplifiers.

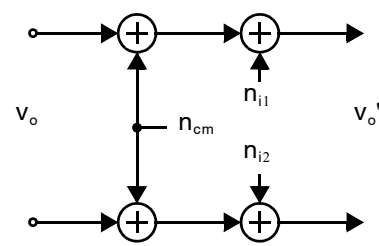


Fig. 6.26 The impact of additive noise on fully differential signals.

therefore rejected by fully differential circuits. In reality, this rejection only partially occurs since, unfortunately, the mechanisms introducing the noise are usually nonlinear with respect to voltage levels. For example, substrate noise will usually feed in through junction capacitances, which are nonlinear with voltage. Also, the clock feedthrough noise introduced by switches turning off usually has some voltage-dependent nonlinearities that can cause more noise to feed into one path than the other, thereby injecting a differential noise. However, almost certainly, the noise rejection of a fully differential design will be much better than that of a single-ended output design.

Key Point: *A significant advantage of fully differential amplifiers is that common mode noise is rejected. Although random noise power may be 2-times greater than in a similar single-ended circuit, the increased signal swing available in fully differential circuits generally affords them a higher maximum signal-to-noise ratio.*

There are also significant random noise sources which appear completely uncorrelated in each of the two half-signals: n_{i1} and n_{i2} in Fig. 6.26.[16] Under the practical assumption that n_{i1} and n_{i2} have the same variance, N_o^2, the noise variance observed at the differential output is the sum of the variances of the two noise contributors, hence $2N_o^2$. This increased random noise may at first appear to be a drawback of fully differential signals. However, since the available voltage swing in fully differential circuits is 2-times greater than in single-ended circuits, the maximum signal power is increased by $2^2 = 4$ -times, resulting in a net 3-dB increase in the maximum achievable signal-to-noise ratio.

Fully differential circuits have the additional benefit that if each single-ended signal is distorted symmetrically around the common-mode voltage, the differential signal will have only odd-order distortion terms (which are often much smaller). To see this distortion improvement, consider the block diagram shown in Fig. 6.27, where the two nonlinear elements are identical and, for simplicity, the common-mode voltage is assumed to be

16. An important practical example of uncorrelated noise is thermal noise, studied in detail in Chapter 9.

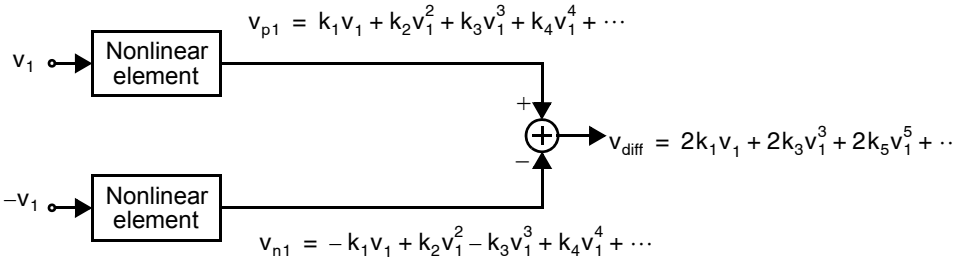

$v_{p1} = k_1v_1 + k_2v_1^2 + k_3v_1^3 + k_4v_1^4 + \cdots$
v_1
Nonlinear element
+
$v_{diff} = 2k_1v_1 + 2k_3v_1^3 + 2k_5v_1^5 + \cdots$
−
$-v_1$
Nonlinear element
$v_{n1} = -k_1v_1 + k_2v_1^2 - k_3v_1^3 + k_4v_1^4 + \cdots$

Fig. 6.27 Demonstrating that even-order distortion terms cancel in fully differential circuits if the distortion is symmetrical around the common-mode voltage. Here, the common-mode voltage is assumed to be zero.

zero. If the nonlinearity is memoryless, then the outputs can be models by the following Taylor series expansion,

$$v_o = k_1v_i + k_2v_i^2 + k_3v_i^3 + k_4v_i^4 + \cdots \tag{6.148}$$

where k_i are constant terms. In this case, the differential output signal, v_{diff}, consists of only the linear term, $2k_1v_1$, and odd-order distortion terms (which are typically smaller than the second-order term). With these two important advantages, most modern analog signal processing circuits are realized using fully differential structures.

One drawback of using fully differential opamps is that a common-mode feedback (CMFB) circuit must be added. This extra circuitry is needed to establish the common-mode (i.e., average) output voltage. Consider the simple fully differential stage in Fig. 6.24(*b*) in the presence of a tiny error in the value of the tail current. Even a 0.0001% increase in the tail current not accompanied by a corresponding increase in the active load currents causes all nodes in the circuit to decrease slowly towards ground, eventually leaving some transistors in triode and degrading the circuit's gain. In practice, this is avoided by a CMFB circuit that somehow observes the output voltages and establishes currents that keep all transistors biased properly. Ideally, the CMFB will keep the common-mode voltages immovable, preferably close to halfway between the power-supply voltages, even when large differential signals are present.

> ***Key Point:*** *Compared to single-ended circuits, fully-differential amplifiers offer reduced harmonic distortion. However, they require additional common-mode feedback circuitry in order to establish the signals' common-mode levels.*

Although in the simple example of Fig. 6.24 the power dissipation is the same for both fully-differential and single-ended output circuits, in general some additional power consumption may be required by fully differential circuits. Specifically, some power may be consumed by the CMFB. In addition, fully differential current mirror opamps and two-stage opamps require additional power consumption in order to produce two output signals instead of one.

6.7.1 Fully Differential Folded-Cascode Opamp

A simplified schematic of a fully differential folded-cascode opamp is shown in Fig. 6.28. The n-channel current-mirror of Fig. 6.20 has been replaced by two cascode current sources, one composed of Q_7 and Q_8, the other composed of Q_9 and Q_{10}. In addition, a common-mode feedback (CMFB) circuit has been added. The gate voltage on the drive transistors of these current sources is determined by the output voltage, V_{cntrl}, of the common-mode feedback circuit. The inputs to the CMFB circuit are the two outputs of the fully differential amplifier. The CMFB circuit will detect the average of these two outputs and force it to be equal to a predetermined value, as we shall see in the next section.

Note that when the opamp output is slewing, the maximum current available for the negative slew rate is limited by the bias currents of Q_7 or Q_9. If the CMFB circuit is very fast, these will be increased dynamically to some degree during slewing, but seldom to the degree of a single-ended output fully differential opamp. For this reason, fully differential folded-cascode opamps are usually designed with the bias currents in the output stage equal to

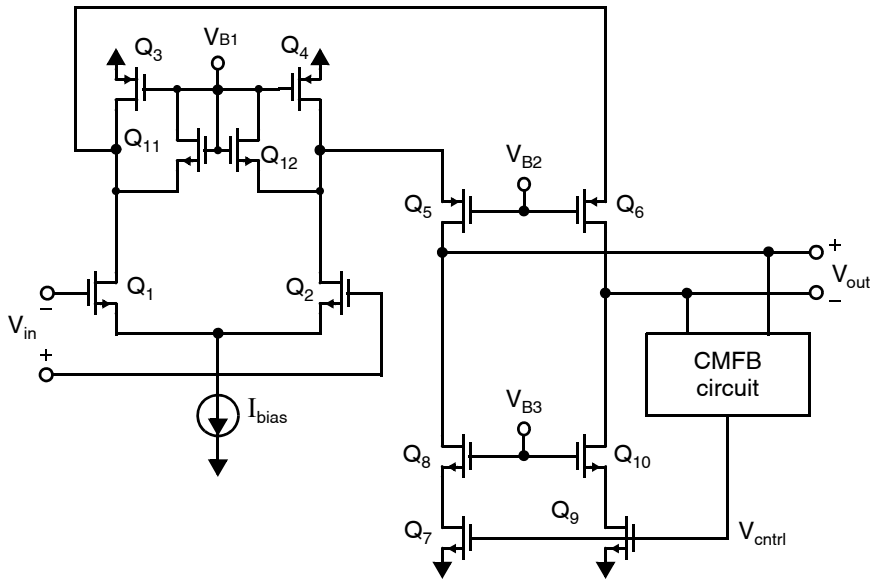


Fig. 6.28 A fully differential folded-cascode opamp.

the bias currents in the input transistors. Also, to minimize transient voltage changes during slew-rate limiting, clamp transistors Q_{11} and Q_{12} have been added, as in the single-ended design presented earlier. If these transistors are not included (as has historically been the case), the time to recover from slew-rate limiting of the folded-cascode amplifier can be comparatively poor.

Each signal path now consists of only one node in addition to the output nodes, namely the drain nodes of the input devices. These nodes will most certainly be responsible for the equivalent second poles. In cases where the load capacitance is small, and it is important to maximize the bandwidth, a complementary design (i.e., n- and p-channel transistors and the power supplies interchanged) would be preferable to maximize the frequency of the equivalent second poles. This complementary circuit would result in the impedance at the drains of the input devices being the inverse of the transconductance of n-channel transistors, rather than p-channel transistors, resulting in smaller impedances and therefore faster time constants. The trade-off involved is that the opamp's transconductance, and therefore the opamp's dc gain, would become smaller due to the input transistors now being p-channel transistors. Also, the CMFB circuitry could possibly be slower, as the current sources being modulated would also have to be p-channel drive transistors. In any case, a complementary design is often a reasonable choice for high-speed fully differential designs.

6.7.2 Alternative Fully Differential Opamps

An alternative design is the fully differential current-mirror opamp, as shown in simplified form in Fig. 6.29. As with the folded-cascode design, this circuit can be realized as shown or in a complementary design, with p-channel input transistors, n-channel current mirrors, and p-channel bias current sources. Which design is preferable is primarily determined by whether the load capacitance or the equivalent second poles are limiting the bandwidth, and by whether maximizing the dc gain or the bandwidth is more important. n-channel inputs are preferable in the former case and p-channel inputs are preferable in the latter case, for reasons similar to those discussed for the folded-cascode opamp. Also, n-channel input transistors will give lower thermal noise (due to their larger transconductances), whereas p-channel inputs will result in less opamp input-referred 1/f noise.

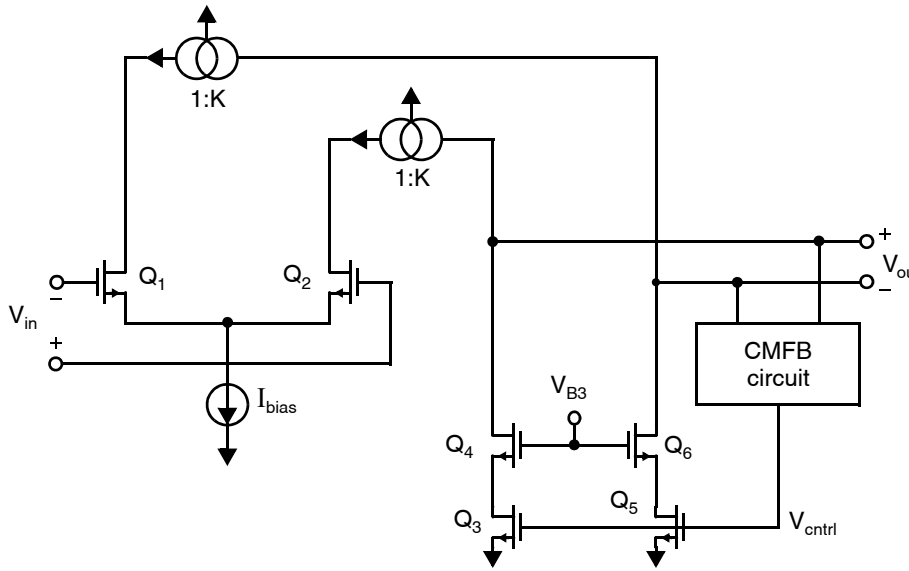


Fig. 6.29 A fully differential current-mirror opamp.

If a general-purpose fully differential opamp is desired, then this design with large p-channel input transistors, a current gain of $K = 2$, and wide-swing enhanced output-impedance cascode mirrors and current sources is probably a good choice, compared to all other fully differential alternatives.

One of the limitations of the fully differential opamps seen so far is that the maximum current at the output for single-ended slewing in one direction is limited by fixed current sources. It is possible to modify the designs to get bidirectional drive capability at the output, as is shown in Fig. 6.30. This opamp is similar to the current mirror design of Fig. 6.29, but now the current mirrors at the top (i.e., the p-channel current mirrors) have been replaced

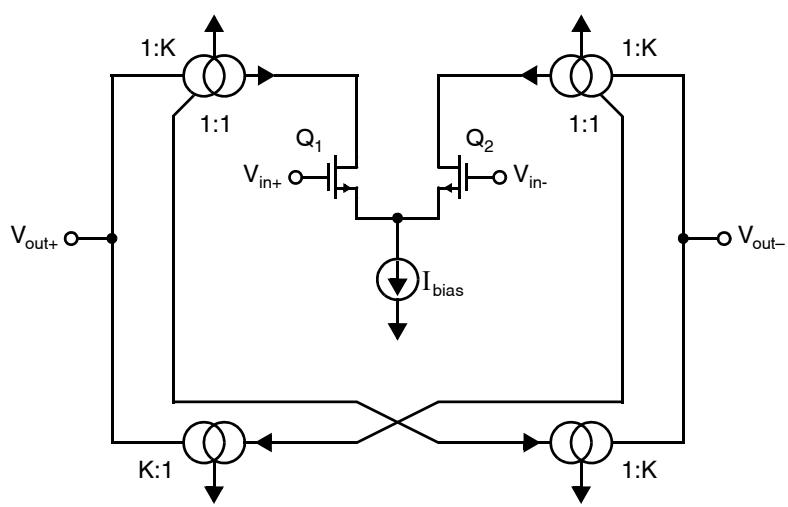


Fig. 6.30 A fully differential opamp with bidirectional output drive.

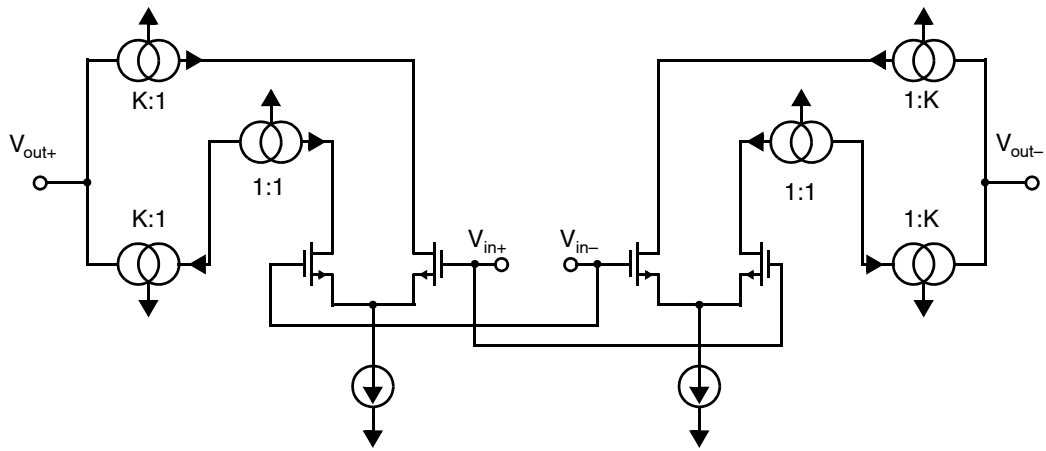


Fig. 6.31 A fully differential opamp composed of two single-ended output current-mirror opamps. CMFB circuit not shown.

by current mirrors having two outputs. The first output has a current gain of K and goes to the output of the opamp, as before. The second output has a gain of one and goes to a new n-channel current mirror that has a current gain of K, where it is mirrored a second time and then goes to the opamp's opposite output, as shown. Assume now the differential output is slewing in the positive direction. For this case, the differential input will be very positive and the slewing current going into V_{out+} will be KI_{bias}, but due to the additional n-channel mirrors, the current being sinked from V_{out-} will also be KI_{bias}. This circuit has an improved slew rate at the expense of slower small-signal response due to the addition of the extra outputs on the p-channel mirrors and the additional n-channel mirrors, although in many applications this trade-off may be well merited. Of course, common-mode feedback must be added to the circuit of Fig. 6.30.

Another alternative for a fully differential opamp is to use two single-ended output opamps with their inputs connected in parallel and each of their outputs being one output side of the fully differential circuit. An example of such an approach using current-mirror subcircuits is shown in Fig. 6.31. This design has a fairly large slew rate when compared to the simpler fully differential current-mirror opamp of Fig. 6.29, but has additional current mirrors and complexity. Another advantage of this design is that the noise voltages due to the input transistors sum in a squared fashion, while the two signal paths sum linearly.[17] As a result, this design has an improvement in signal-to-noise ratio of the input-referred gain by a factor of $\sqrt{2}$, or 3 dB. Also, the increase in total power dissipation is not significant when K is moderately large. As for the previous design, the compensation of the common-mode feedback loop is more difficult than for designs having fixed current sources biasing the output stages.

6.7.3 Low Supply Voltage Opamps

Low supply voltages complicate opamp design considerably. The input common-mode voltage must be restricted within a very tight range to ensure the input differential pair tail current remains in active mode. For example, referring to the folded cascode opamp in Fig. 6.28, and assuming that a simple NMOS transistor is used as the tail current source I_{bias}, the input common-mode voltage must be greater than $V_{GS1} + V_{eff}$ in order to keep the tail

17. See Chapter 4 regarding noise.

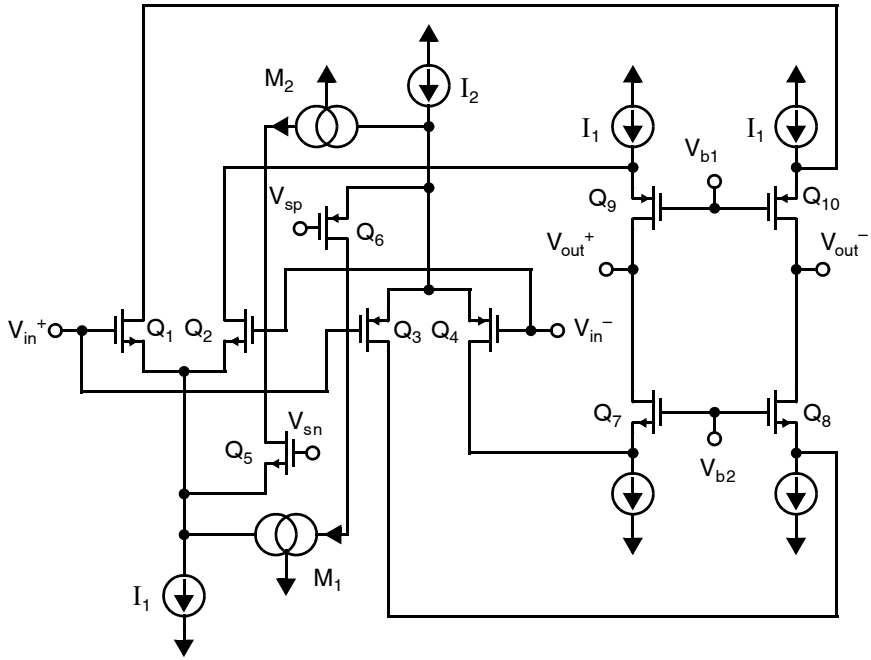


Fig. 6.32 An opamp having rail-to-rail input common-mode voltage range. CMFB circuit not shown.

current source device in active mode. Assuming $V_{eff} \cong 200$ mV, $V_{tn} \cong 450$ mV, and an additional 50 mV of margin to account for process and temperature variations, this means the input common model voltage must be at least 950 mV , a significant limitation if the supply voltage is only 1.2 V.

> ***Key Point:*** *Low supply-voltages dictate the use of special circuit techniques to permit common-mode signal levels near the supply rails and/or to provide sufficient gain and output swing.*

The opamp shown in somewhat simplified form in Fig. 6.32 makes use of both n-channel and p-channel transistors in the two differential input pairs. Hence, the input common-mode voltage range is increased [Babanezhad, 1988; Hogervorst, 1992; Coban, 1994], a particularly important feature when low power-supply voltages are being used. When the input common-mode voltage range is close to one of the power-supply voltages, one of the input differential pairs will turn off, but the other one will remain active. In an effort to keep the opamp gain relatively constant during this time, the bias currents of the still-active differential pair are dynamically increased. For example, if the input common-mode voltage range was close to the positive power-supply voltage, Q_3 and Q_4 would turn off, and Q_6 would conduct all of I_2. This current would go through current mirror M_1 and increase the bias current of the differential pair consisting of Q_1 and Q_2, which is still active. A similar situation occurs if the input common-mode voltage is near the negative power-supply rail. With careful design, it has been reported that the transconductance of the input stage can be held constant to within 15 percent of its nominal value with an input common-mode voltage range as large as the difference between the power-supply voltages [Coban, 1994].

Another challenge of low supply voltage designs is that it is very difficult to use single-stage folded-cascode architectures while maintaining a reasonable output swing. Referring to the folded-cascode schematic in Fig. 6.28, notice that the output nodes must remain at least $2V_{eff}$ away from both the positive supply and ground in order to keep all of transistors Q_3 – Q_{10} in active mode. For example, if a 1.2-V supply is to be used, assuming $V_{eff} \approx 200$ mV, and allowing a additional 50 mV of margin on both the NMOS and PMOS cascode voltages, the

output nodes will only be able to swing within the range 450 mV to 750 mV. The opamp circuit in Fig. 6.33 [Dessouky, 2001] is a two-stage opamp with a folded cascode first stage and common-source second stage. The cascode stage can provide high gain, but is not required to drive the opamp outputs, and hence is not required to accommodate the full output swing. The signal swing at V_1^+ and V_1^- will be less than the output swing by a factor equal to the small-signal gain of the common-source second stage. Miller compensation is provided, just as with the conventional two-stage opamp, suing C_C and R_C. This architecture can provide large gain even within a low supply voltage.

6.8 COMMON-MODE FEEDBACK CIRCUITS

Typically, when using fully-differential opamps in a feedback application, the applied feedback determines the differential signal voltages, but does not affect the common-mode voltages. It is therefore necessary to add additional circuitry to determine the output common-mode voltage and to control it to be equal to some specified voltage, usually about halfway between the power-supply voltages. This circuitry, referred to as the *common-mode feedback* (CMFB) *circuitry,* is often the most difficult part of the opamp to design.

There are two typical approaches to designing CMFB circuits—a continuous-time approach and a switched-capacitor approach. The former approach is often the limiting factor on maximizing the signal swings, and, if nonlinear, may actually introduce common-mode signals. The latter approach is typically only used in switched-capacitor circuits, since in continuous-time applications it introduces clock-feedthrough glitches.

An example of a continuous-time CMFB circuit is shown in Fig. 6.34 [Martin, 1985; Whatly, 1986]. To illustrate its operation, assume a common-mode output voltage, $V_{out,CM}$, equal to the reference voltage, $V_{ref,CM}$, and that V_{out+} is equal in magnitude, but opposite in sign, to V_{out-}. Furthermore, assume the two differential pairs have infinite common-mode input rejection, which implies that the large-signal output currents of the differential pairs depend only on their input differential voltages. Since the two pairs have the same differential voltages being applied, the current in Q_1 will be equal to the current in Q_3, while the current in Q_2 will equal the current in Q_4. This result is valid independent of the nonlinear relationship between a differential pair's input voltage and its large-signal differential drain currents. Now, letting the current in Q_2 be denoted as $I_{D2} = I_B/2 + \Delta I$, where I_B is

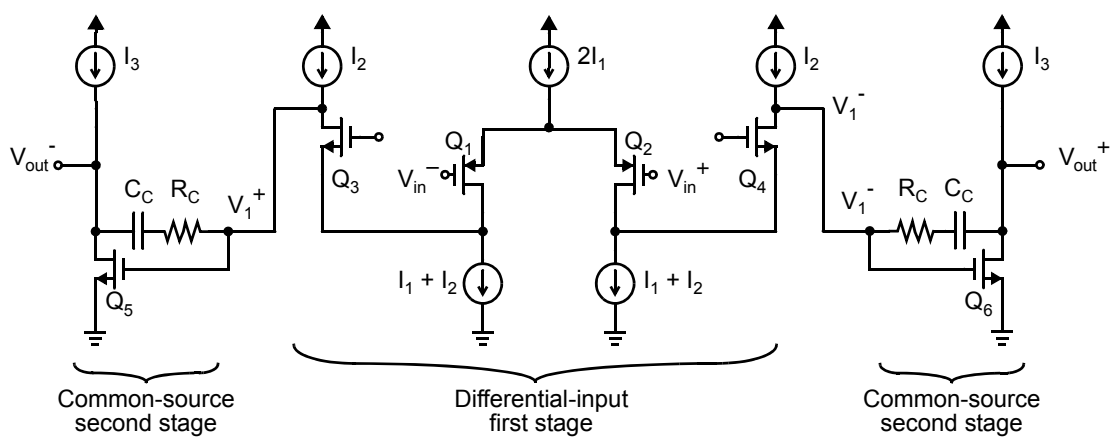


Fig. 6.33 A differential Miller-compensated two-stage amplifier with a folded cascode first-stage [Dessouky, 2001]. (Common-mode feedback not shown.)

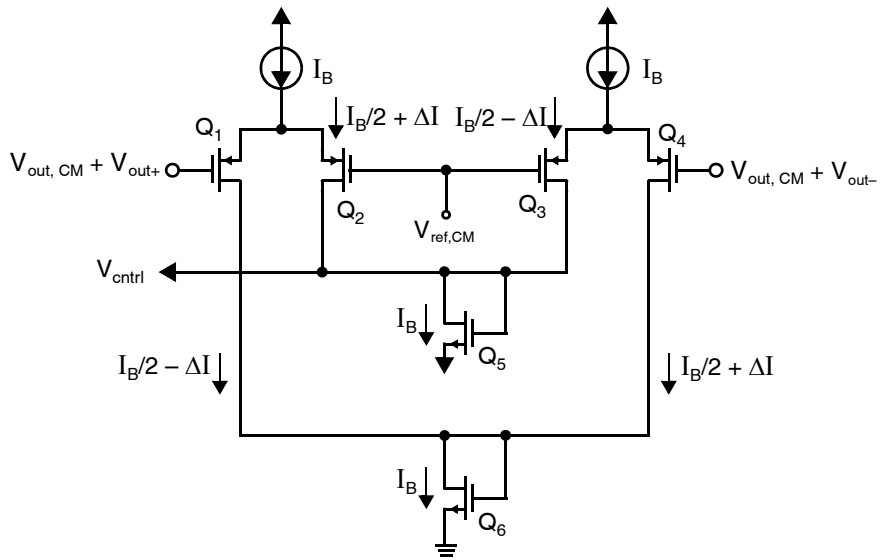


Fig. 6.34 An example of a continuous-time CMFB circuit.

the bias current of the differential pair and ΔI is the large-signal current change in I_{D2}, the current in Q_3 is given by $I_{D3} = (I_B/2) - \Delta I$, and the current in Q_5 is given by

$$I_{D5} = I_{D2} + I_{D3} = (I_B/2 + \Delta I) + ((I_B/2) - \Delta I) = I_B \tag{6.149}$$

Thus, as long as the voltage V_{out+} is equal to the negative value of V_{out-}, the current through diode-connected Q_5 will not change even when large differential signal voltages are present. Since the voltage across diode-connected Q_5 is used to control the bias voltages of the output stage of the opamps, this means that when no common-mode voltage is present, the bias currents in the output stage will be the same regardless of whether a signal is present or not. Note, however, that the above result does not remain valid if the output voltage is so large that transistors in the differential pairs turn off.

Next consider what happens when a common-mode voltage greater than $V_{CM,\,ref}$ is present. This positive voltage will cause the currents in both Q_2 and Q_3 to increase, which causes the current in diode-connected Q_5 to increase, which in turn causes its voltage to increase. This voltage is the bias voltage that sets the current levels in the n-channel current sources at the output of the opamp (e.g. transistors $Q_{7,9}$ in Fig. 6.28 or $Q_{3,5}$ in Fig. 6.29). Thus, both current sources will have larger currents pulling down to the negative rail, which will cause the common-mode output voltage, $V_{out,\,CM}$, to decrease towards $V_{ref,\,CM}$. Thus, as long as the common-mode loop gain is large enough, and the differential signals are not so large as to cause transistors in the differential pairs to turn off, the common-mode output voltage will be kept very close to ground $V_{ref,\,CM}$.

The size of the differential signals that can be processed without one of the differential-pair signals turning off is maximized if the differential-pair transistors are designed to have large effective gate-source voltages. Additionally, source degeneration can be used to allow them to have larger input signals without all of the current being directed to one side of the differential pair, as shown in Fig. 6.35. However, even when this maximization is performed, the CMFB circuit may still limit the differential signals to be less than what can be processed by the rest of the opamp.

Finally the current sources I_B should be high-output impedance cascode current sources to ensure good common-mode rejection of the two differential pairs.

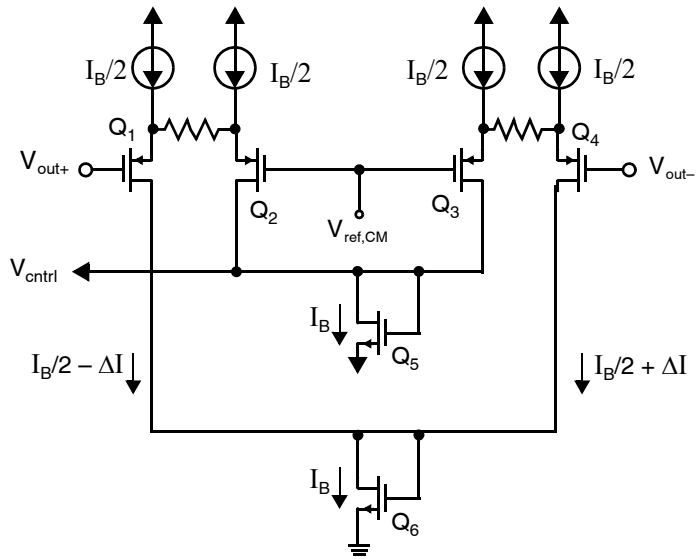


Fig. 6.35 A continuous-time CMFB circuit that can accommodate increased output swing.

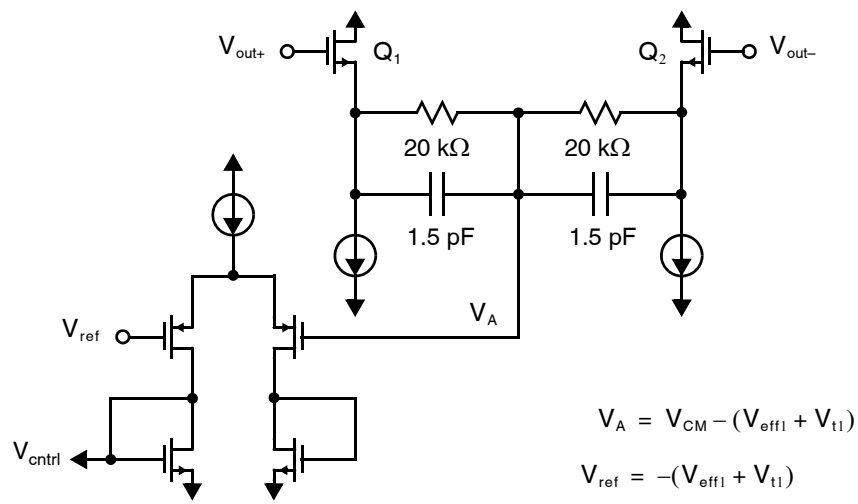


Fig. 6.36 An alternative continuous-time CMFB circuit.

An alternative approach for realizing CMFB circuits is shown in Fig. 6.36 [Banu, 1988]. This circuit generates the common-mode voltage of the output signals (minus a dc level shift) at node V_A. This voltage is then compared to a reference voltage, V_{ref}, using a separate amplifier. Although this approach works well, it has a major limitation in that the voltage drop across the source-follower transistors Q_1, Q_2 severely limits the differential signals that can be processed (unless transistors with native threshold voltages, such as 0.3 volts, are available). This limitation is particularly important when small power-supply voltages are used. In addition, the additional nodes of the common-mode feedback circuitry make the circuit slightly more difficult to compensate. When bipolar transistors are available, as is the case in a BiCMOS process, this approach is much more desirable.

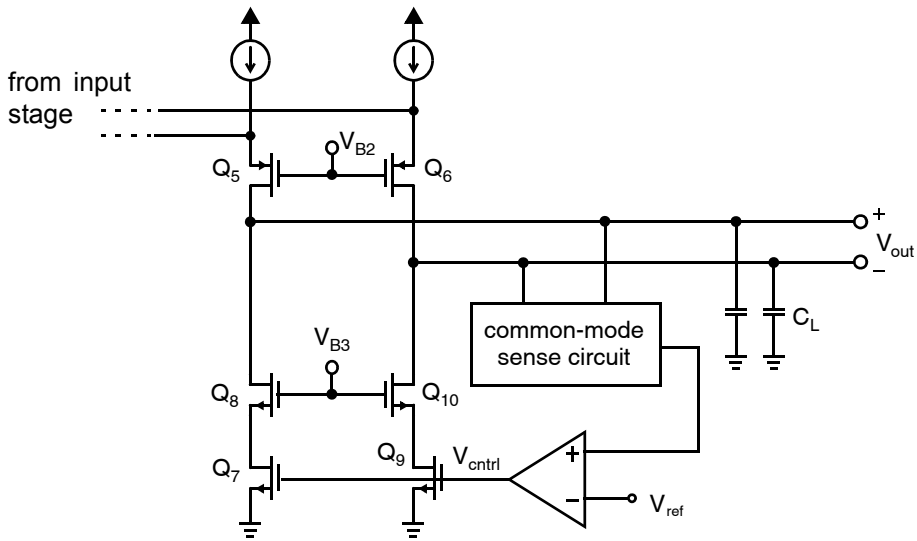


Fig. 6.37 The common-mode feedback loop in a fully differential folded-cascode opamp.

An important consideration when designing CMFB circuits is that they are part of a negative feedback loop, as illustrated in Fig. 6.37, and must therefore be well compensated, otherwise the injection of common-mode signals can cause them to ring or even possibly become unstable. Thus, when the circuit is being designed, the phase margin and step-response of the common-mode loop should be found and verified by simulation. Phase margin might be observed, for example, by breaking the gate connections at V_{cntrl}, applying a test small-signal and looking at the returned signal. A step-response test could be performed by applying a small step to the common-mode reference signal, $V_{ref,CM}$.

Compensation of the CMFB loop may be difficult because it includes two high-impedance nodes: V_{cntrl} and the amplifier output. Moreover, high bandwidth is desirable in order to suppress high-frequency common-mode noise that may appear due to neighbouring on-chip clock signals. However, it is not generally necessary to set the output common-mode level extremely precisely; a few millivolts of error should certainly be tolerable. Hence, very high dc gain is not required in the CMFB. In fact too much dc loop gain may complicate compensation.

Often, the common-mode loop is stabilized using the same capacitors used to compensate the differential loop. This multipurpose compensation is achieved by connecting two compensation (or loading) capacitors between the opamp outputs and ground (or some other reference voltage). However, if the deferential loop is compensated using a single compensation capacitor connected directly between the two outputs, the common-mode loop remains uncompensated. It should be mentioned that by having as few nodes in the common-mode loop as possible, compensation is simplified without having to severely limit the speed of the CMFB circuit. For this reason, the CMFB circuit is usually used to control current sources in the output stage of the opamp, as opposed to the current sources in the input stage of the opamp. For the same reason, in two-stage fully-differential opamps such as in Fig. 6.33, the common-mode output voltage of each stage is observed and controlled using a separate CMFB circuit. High speed in the CMFB circuit is necessary to minimize the effects of high-frequency common-mode noise, which could otherwise be amplified, causing the opamp outputs to saturate. Designing continuous-time CMFB circuits that are both linear and operate with low power-supply voltages is an area of continuing research.

A third approach for realizing CMFB circuits is based on the use of switched-capacitor circuits. An example of this approach is shown in Fig. 6.38 [Senderowicz, 1982; Castello, 1985]. In this approach, capacitors labelled C_C generate the average of the output voltages, which is used to create control voltages for the opamp current

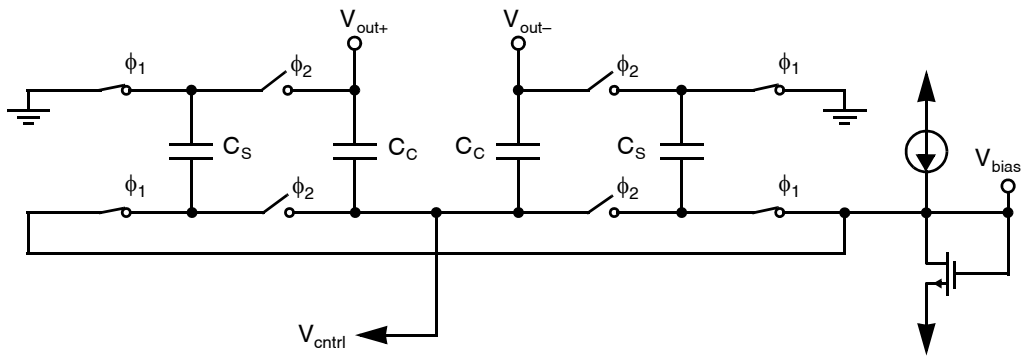


Fig. 6.38 A switched-capacitor CMFB circuit.

sources. The dc voltage across C_C is determined by capacitors C_S, which are switched between bias voltages and between being in parallel with C_C. This circuit acts much like a simple switched-capacitor low-pass filter having a dc input signal. The bias voltages are designed to be equal to the difference between the desired common-mode voltage and the desired control voltage used for the opamp current sources.

Key Point: *Common-mode feedback may be performed using switched-capacitor circuits in discrete-time applications. When continuous-time signals are present, the design of a suitable CMFB can be a significant challenge because it introduces constraints on output swing, and may reduce the opamp's load impedance.*

The capacitors being switched, C_S, might be between one-quarter and one-tenth the sizes of the fixed (not switched) capacitors, C_C. Using larger capacitance values overloads the opamp more than is necessary during the phase ϕ_2, and their size is not critical to circuit performance. Reducing the capacitors too much causes common-mode offset voltages due to charge injection of the switches. Normally, all of the switches would be realized by minimum-size n-channel transistors only, except for the switches connected to the outputs, which might be realized by transmission gates (i.e., parallel n-channel and p-channel transistors both having minimum size) to accommodate a wider signal swing.

In applications where the opamp is being used to realize switched-capacitor circuits, switched-capacitor CMFB circuits are generally preferred over their continuous-time counterparts since they allow a larger output signal swing.

6.9 SUMMARY OF KEY POINTS

- The classic two-stage opamp comprises a differential input gain stage, a common-source second gain stage, and optionally a unity-gain output buffer stage. [p. 243]
- The dc gain of the two-stage opamp is approximated by the product of two stages, each having a gain of approximately $g_m r_{ds}/2$. [p. 244]
- The two-stage opamp's frequency response is dominated by a pole at the input to the second stage arising from the introduced Miller capacitor, C_C. A dominant-pole approximation yields a unity-gain frequency of g_{m1}/C_C. [p. 246]
- The two-stage opamp's second pole arises at the output node and may be increased by increasing the second-stage transconductance, gm_7. [p. 248]
- The slew rate is the maximum rate at which the output of an opamp changes when a large differential input signal is present. [p. 249]

- The only ways of improving the slew rate of a two-stage opamp is to increase V_{eff1} or ω_{p2}. [p. 251]
- When using a two-stage opamp, p-channel input transistors are almost always the best choice for the first stage because they offer larger unity-gain frequency and lower 1/f noise, with the major disadvantage being an increase in wideband thermal noise. [p. 252]
- In general, dominant-pole compensation is performed by decreasing the frequency of one pole in a feedback loop while leaving the others relatively unchanged. [p. 254]
- Lead compensation in opamp circuits introduces into the loop response a left half-plane zero at a frequency slightly greater than the unity gain frequency to provide a small phase lead and, therefore, and increase in phase margin. [p. 255]
- The feedback network must be accounted for in compensation. [p. 255]
- Compensation of the two-stage opamp proceeds by selecting Miller capacitor C_C to provide 55 degrees phase margin, then introducing a lead compensation series resistor to add another 30 degrees of phase margin. [p. 257]
- Cascoding is a key technique to enable high gain using transistors with low drain-source resistance, but if the cascode transistor is improperly biased, output swing is greatly reduced. [p. 262]
- Auxiliary amplifiers may be used to enhance the output impedance of cascode current mirrors and gain stages, thereby boosting dc gain. [p. 264]
- Operational transconductance amplifiers provide a large dc gain from a single stage. Their high output impedance makes them suitable for driving the mostly-capacitive loads encountered in CMOS integrated circuits. With feedback, they can also drive some resistive load. [p. 268]
- The folded-cascode opamp is a transconductance amplifier whose frequency response is dominated by the output pole at $1/R_{out}C_L$. [p. 269]
- If the bandwidth of a folded-cascode opamp is insufficient, one must increase the input transconductance. [p. 270]
- If the phase margin of a folded-cascode opamp is insufficient, one must either reduce the bandwidth by adding output capacitance, or burn more power in the output transistors to increase the second pole frequency. [p. 271]
- For a current-mirror opamp, when the load capacitance is large and the output pole clearly dominant, a large current mirror ratio, K, maximizes gain and bandwidth. If the second pole frequency becomes significant, lower values of K are required to keep the second pole frequency above the unity-gain frequency and maintain a healthy phase margin. [p. 277]
- Fully differential amplifiers generally have similar power consumption, but twice the available signal swing of their single-ended counterparts. They can also exhibit similar small signal gain and bandwidth. [p. 281]
- A significant advantage of fully differential amplifiers is that common mode noise is rejected. Although random noise power may be 2-times greater than in a similar single-ended circuit, the increased signal swing available in fully differential circuits generally affords them a higher maximum signal-to-noise ratio. [p. 282]
- Compared to singleended circuits, fully-differential amplifiers offer reduced harmonic distortion. However, they require additional commonmode feedback circuitry in order to establish the signals' common-mode levels. [p. 283]
- Low supply-voltages dictate the use of special circuit techniques to permit common-mode signal levels near the supply rails and/or to provide sufficient gain and output swing. [p. 287]
- Common-mode feedback may be performed using switched-capacitor circuits in discrete-time applications. When continuous-time signals are present, the design of a suitable CMFB can be a significant challenge because it introduces constraints on output swing, and may reduce the opamp's load impedance. [p. 292]

6.10 REFERENCES

J. N. Babanezhad. "A Rail-to-Rail CMOS Opamp," *IEEE J. of Solid-State Circuits,* Vol. 23, no. 6, pp. 1414–1417, December 1988.

J. N. Babanezhad and R. Gregorian. "A Programmable Gain/Loss Circuit," *IE EE J. of Solid-State Cir cuits,* Vol. 22, no. 6, pp. 1082–1090, December 1987.

M. Banu, J. M. Khoury, and Y. Tsividis. "Fully Differential Operational Amplifiers with Accurate Output Balancing," *IEEE J. of Solid-State Circuits,* Vol. 23, no. 6, pp. 1410–1414, December 1988.

K. Bult and G. J. G. M. Geelen. "A Fast-Settling CMOS Opamp for SC Circuits with 90-dB DC Gain," *IEEE J. of Solid-State Circuits,* Vol. 25, no. 6, pp. 1379–1384, December 1990.

A. Coban and P. Allen. "A 1.75-V Rail-to-Rail CMOS Opamp," *Proceedings of the IEEE Int. Symp. on Circuits and Systems,* Vol. 5, pp. 5.497–5.500, London, June 1994.

M. Dessouky, A. Kaiser. "Very Low-Voltage Digital-Audio DS Modulator with 88-dB Dynamic Range Using Local Switch Bootstrapping," *IEEE J. of Solid-State Circuits,* Vol. 36, no. 3, pp. 349–355, March 2001.

U. Gatti, F. Maloberti, and G. Torelli. "A Novel CMOS Linear Transconductance Cell for Continuous-Time Filters," proceedings of the *IEEE Int. Symp. on Circuits and Systems,* pp. 1173–1176, New Orleans, May 1990.

P. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits,* 5th ed. John Wiley & Sons, New York, 2009.

R. Hogervorst, et al. "CMOS Low-Voltage Operational Amplifiers with Constant-G_m Rail-to-Rail Input Stage," proceedings of the *IEEE Int. Symp. on Circuits and Systems,* pp. 2876–2879, San Diego, May 1992.

B. Hosticka. "Improvement of the Gain of MOS Amplifiers," *IEEE J. of Solid-State Circuits,* Vol. SC-14, no. 14, pp. 1111–1114, December 1979.

S. Law. Private conversation, Xerox Corp., 1983.

K. Martin. Class notes, UCLA, 1985.

K. Martin. Laboratory notes (independently derived, albeit after [Coban 1994]), 1994.

J. K. Roberge. *Operational Amplifiers*. John Wiley & Sons, New York, 1975.

E. Säckinger and W. Guggenbühl. "A High-Swing, High-Impedance MOS Cascode Circuit," *IEEE J. of Solid-State Circuits,* Vol. 25, no. 1, pp. 289–298, February 1990.

A. S. Sedra and K. C. Smith. *Microelectronic Circuits*, 6th ed. Oxford University Press, New York, 2010.

D. Senderowicz, et al. "A Family of Differential NMOS Analog Circuits for a PCM Codec Filter Chip," *IEEE J. of Solid-State Circuits,* Vol. 17, no. 6, pp. 1014–1023, December 1982.

C. C. Shih and P. R. Gray. "Reference Refreshing Cyclic Analog-to-Digital and Digital-to-Analog Converters," *IEEE J. of Solid-State Circuits,* Vol. 21, no. 4, pp. 544–554, August 1986.

N. S. Sooch. "MOS Cascode Current Mirror," U.S. patent no. 4,550,284, October 1985.

J. M. Steininger. "Understanding Wide-Band MOS Transistors," *IEEE Circuits and Devices*, Vol. 6, No. 3, pp. 26–31, May 1990.

R. A. Whatly. "Fully Differential Operational Amplifier with DC Common-Mode Feedback," U.S. patent no. 4573020, February 1986.

6.11 PROBLEMS

Unless otherwise stated, assume the process parameters for the 0.18-μm process in Table 1.5.

SPICE! *Most of these problems can be performed with the aid of SPICE simulations using the provided device models.*

6.11.1 Section 6.1: Two-Stage CMOS Opamp

6.1 For the two-stage opamp of Fig. 6.3, assume that the input bias current is reduced to $I_{D8} = 15\ \mu A$, all transistor sizes remain the same, and $C_c = 2$ pF.

- **a.** Estimate the dc gain of the opamp.
- **b.** Estimate the –3-dB frequency of the first stage.
- **c.** Estimate the unity-gain frequency of the opamp.
- **d.** Find the slew rate.

6.2 If all of the gate lengths in Fig. 6.3 are increased to 0.4 μm, what is the resulting impact on

- **a.** the dc gain of the opamp
- **b.** the unity-gain frequency of the opamp

6.3 In the two-stage opamp of Fig. 6.3, $C_C = 0.5$ pF.

a. What is the slew rate?

b. What circuit changes could be made to increase the slew rate by 30% while keeping ω_t and C_C unchanged?

c. What circuit changes could be made to increase the slew rate by 20% while keeping ω_t and power consumption unchanged?

6.4 How should the sizes of Q_6 and Q_7 change in Problem 6.3 to maintain zero systematic offset voltage?

6.5 Assume Q_7 of the two-stage opamp in Fig. 6.3 is widened to 21 μm wide, and everything else remains the same. Estimate the inherent input-offset voltage.

6.6 Ignoring the body effect, what is the output voltage range of the opamp of Fig. 6.3? What is the range of acceptable common-mode input voltage, assuming a 1.8-V power supply is used?

6.7 Repeat Problem 6.6, but take the body effect into account and find approximate answers. Assume the body terminals of $Q_{1,2}$ are connected to V_{DD}, $\Phi f \approx 0.4$ and $\gamma = 0.3\ V^{-1/2}$

6.8 Assume the two-stage opamp of Fig. 6.3 is connected with a resistive feedback network, as shown in Fig. P6.8. If $V_{in} = 0.5$ V, how much current flows through R_2? Hence, what is the remaining current flowing through Q_7? Estimate the effect of this on the transconductance g_{m7} and the second pole frequency, ω_{p2}.

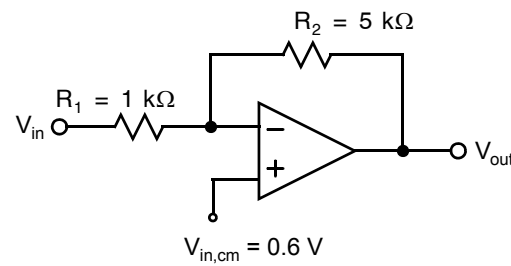


Fig. P6.8

6.11.2 Section 6.2: Opamp Compensation

6.9 A two-stage opamp with compensation capacitor has an equivalent second-pole frequency of 60 MHz. Assume the input transistors of the first stage have a transconductance of 3 mA/V and that there is an output buffer with a gain of exactly unity. What is the required size of the compensation capacitor if the phase margin is to be 55° for the feedback configuration shown in Fig. P6.8?

6.10 For the result of Problem 6.9, what is a near-optimum size for a capacitor to be placed in parallel with R_2 in order to obtain lead compensation?

6.11 You are to compensate the two-stage opamp of Fig. 6.3 when driving a load capacitance of $C_L = 5$ pF.

a. Following the procedure in section 6.2.2, select the values of C_C and R_C in the compensation network assuming a unity-gain feedback configuration.

b. Keeping the same values of C_C and R_C as in part (a), what is the maximum load capacitance that can be driven while maintaining at least 60° phase margin?

c. Modify C_C, R_C, and $W_{1,2}$ to increase the slew rate by 30% while maintaining the same phase margin and total power consumption. Estimate the resulting impact on the opamp's dc gain.

6.12 Assume a two-stage opamp has been optimally compensated for a noninverting closed-loop gain of 4 (i.e., $\beta = 0.25$) using the procedure in section 6.2.2 so that the equivalent second pole is 43% above the unity-gain frequency, $\omega_{eq} = 1.43\omega_t$, and the lead compensation zero is 70% above the unity-gain frequency, $\omega_z = 1.7\omega_t$. What is the phase margin if the feedback network is modified to provide a closed-loop gain of 2 (i.e., $\beta = 0.5$)? What about a unity-gain configuration ($\beta = 1$)?

6.13 It is desired to increase the width, and hence drain currents, of $Q_{6,7}$ in the two-stage opamp of Fig. 6.3 in order to improve the amplifier's bandwidth. Assuming a load capacitance of 1 pF, estimate the width of $Q_{6,7}$ required to provide a second pole frequency of 300 MHz.

6.14 Consider what happens when Q_9 is placed on the output side of C_C in Fig. 6.10 (rather than where it is shown). Explain why the resulting two-stage opamp circuit would then oscillate for large positive output values.

6.15 An opamp has an open-loop transfer function given by

$$A(s) = \frac{A_0\left(1 + \frac{s}{\omega_z}\right)}{\left(1 + \frac{s}{\omega_1}\right)\left(1 + \frac{s}{\omega_2}\right)}$$

Assume $A_0 = 10^4$ and $\omega_2 = 10^8$ rad/s. The feedback network has a gain $\beta = 0.5$. The frequency of the zero is 70% higher than the resulting open-loop unity-gain frequency, $\omega_z = 1.7\omega_t$. Find ω_1 and ω_t so that the phase margin is 80°.

6.16 An opamp has an open-loop transfer function given by

$$A(s) \cong \frac{A_0(1 + s\tau_z)}{s\tau_1(1 + s\tau_2)}$$

Find the transfer function of the closed-loop amplifier, assuming a feedback factor β exists. Find approximate equations for the resonant frequency and the Q factor of the denominator of the transfer function of the closed-loop amplifier.

6.17 Prove that the frequencies of the first and second poles of the amplifier of Fig. 6.3 are still given by (6.19) and (6.20), even when lead compensation is used (i.e., when R_C is included).

6.11.3 Section 6.3: Advanced Current Mirrors

6.18 Calculate the output impedance of the two-transistor diode-connected circuit shown in Fig. P6.18 using small-signal analysis. Assume both transistors are in the active region, ignore the body effect, and assume $g_{m1} = g_{m2}$, $r_{ds1} = r_{ds2}$, and $g_m r_{ds} >> 1$.

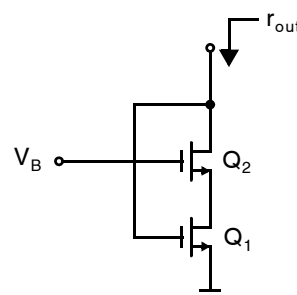


Fig. P6.18

6.19 Assume n is chosen to be 1 in Fig. 6.12 and all transistors are taken to be equal sizes, except $(W/L)_5$ is chosen so that $V_{DS3} = V_{DS2} = V_{eff3} + 0.1$ V, where V_{eff} is chosen to be 0.2 V for all transistors except Q_5. Ignoring the body effect, find the required size for all transistors assuming they all have lengths equal to 0.3 μm. Assume that $I_{bias} = 50\ \mu A$.

6.20 For the circuit in Problem 3.23 (page 147) replace Q_5 and Q_6 with a single diode-connected NMOS transistor Q_7 to generate the gate voltage of Q_2. Find the size of Q_7 to maximize the signal swing that can be obtained at the amplifier output, v_o, while ensuring all transistors remain in active mode.

6.21 What are the drain-source voltages of Q_3 and Q_4 from Problem 6.19 if the lengths of Q_3 and Q_4 are decreased to 0.2 μm to maximize speed?

6.22 Using small-signal analysis and ignoring the body effect, show that the circuit of Fig. 6.13 has an output impedance given by $r_{out} \cong g_{m1} r_{ds1} r_{ds2}(1 + A)$.

6.23 A wide-swing cascode current mirror is shown in Fig. P6.23. Assuming all devices are identically sized (W/L), derive an expression for the resistance R_1 in terms of the current I_{in} and device parameters in order to ensure that Q_1 and Q_3 are in active mode. How should R_1 be chosen to provide 50 mV of margin so that

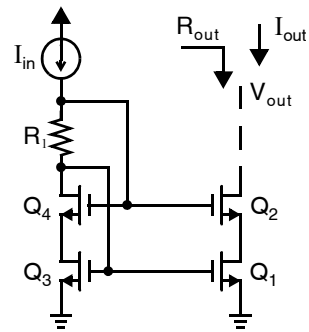


Fig. P6.23

$V_{DS,1} = V_{eff,1} + 50$ mV?

6.24 Consider the compound transistor in Fig. P6.24. Transistor Q_5 is in triode and Q_6 is in active mode. Find the relationship between V_x and V_G in terms of K where $(W/L)_5 = K(W/L)_6$.

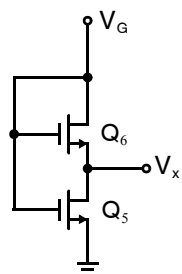


Fig. P6.24

6.25 The resistor R_1 in Fig. P6.23 can be replaced by the triode transistor Q_5 in the circuit of Fig. P6.24. The result is the circuit shown in Fig. P6.25 [Sooch, 1985]. All devices are sized identically except for Q_5 which is sized smaller by a factor $K < 1$. Find the value of K required to keep $Q_{1,3}$ in active mode.

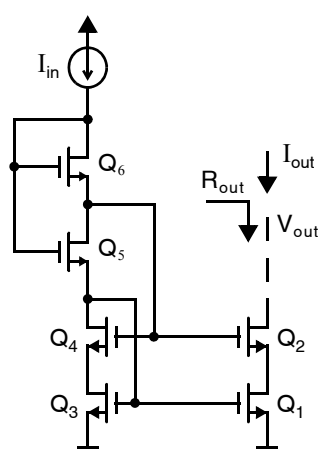


Fig. P6.25

6.26 For the circuit shown in Fig. P6.26, using small-signal analysis, find the output impedance. Approximate this expression for large A and compare this to the approximate expression without the amplifier (i.e., $A = 0$).

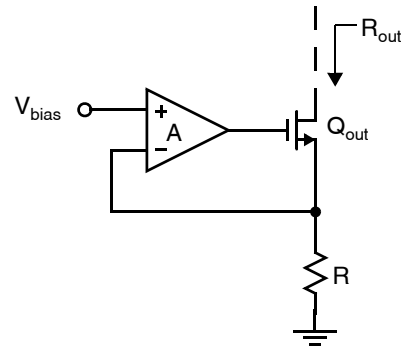


Fig. P6.26

6.27 For the circuit shown in Fig. 6.16 with the W/L ratios shown and assuming all lengths are 0.3 μm, estimate the output impedance. Assume that all current sources are ideal and that $I_{in} = 7I_{bias}$. Compare this to the output impedance of a wide-swing cascode current mirror where the gate of Q_1 was simply connected to a dc bias voltage that keeps Q_2 in active mode.

6.28 The folded cascode amplifier in Fig. P6.28, is to be designed using the process parameters for the 0.35-μm process in Table 1.5. Assume the bias voltage of the input is such that the bias voltage of the output is 1.5 V.

a. Find the device widths so that $V_{eff} = 0.3$ V for all transistors except Q_6 and Q_{11}. Take $V_{eff\text{-}6} = V_{eff\text{-}11} = 2.24V_{eff}$. Take $I_D = 0.25$ mA for all transistors.

b. Find the bias voltages at all nodes and verfify that all transistors are in the active region. (Ignore the body-effect when calculating the nodal voltages.)

c. Find the impedance separately looking into the drain of Q_2 and looking into the drain of Q_4 and also the amplifier output impedance.

d. Find the amplifier gain.

e. Assuming the load capacitance is 0.25pF and dominates, what is the –3dB bandwidth of the the amplifier? What is the unity-gain frequency of the amplifier?

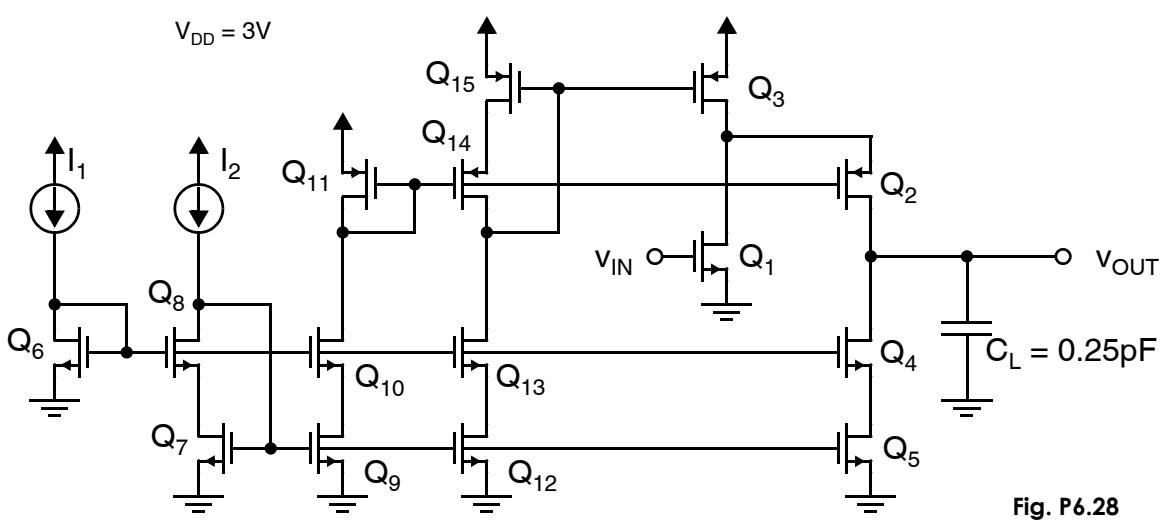


Fig. P6.28

6.11.4 Section 6.4: Folded-Cascode Opamp

6.29 An OTA has $G_{ma} = 5$ mA/V and a dc gain of 45 dB. What is its output resistance, r_{out}?

6.30 An OTA has a unity gain frequency of 50 MHz and a dc gain of 65 dB. What is its dominant pole frequency?

6.31 A folded cascode opamp is to be used in the feedback configuration of Fig. 6.23 with $C_1 = 4$ pF, $C_2 = 2$ pF, and C_p, C_c, C_{load} all negligible. Assume the input differential pair transistors are to be biased with $V_{eff,1} = 150$ mV, and 80% of the total dc current dissipation is to flow through the input differential pair transistors. Estimate the total current consumption required to obtain a closed-loop 3-dB bandwidth of 60 MHz.

6.32 Augment the design of the folded-cascode opamp in Example 6.9 to replace I_{bias} with the wide-swing cascode bias circuit of Fig. 6.12 where $n = 1$ and $I_{in} = I_{bias} = 20$ μA. Next, replace the tail current source I_{bias2} with a NMOS transistor biased by the same wide-swing bias circuit. Finally, introduce additional circuitry to generate the required bias voltage V_{B1}. Give the sizes of all transistors.

6.33 For the folded-cascode amplifier shown in Fig. 6.20, with the transistor sizes given in Table 6.1, $I_{bias,1} = 20$ μA, and $C_L = 10$ pF , find the unity-gain frequency and the slew rate.

6.34 For the folded-cascode amplifier shown in Fig. 6.20, with the transistor sizes given in Table 6.1, estimate the approximate frequency of the second pole caused by the parasitic capacitances at the drains of Q_1 and Q_2. Assume a 1.8-V power supply is used, and the total supply current is 0.4 mA as in Example 6.9. To simplify matters, junction capacitances can be ignored. What should C_L be to achieve 70° phase margin? What would the corresponding unity-gain frequency and slew rate be?

6.35 For the amplifier of Problem 6.34, find the load capacitance, C_L, for a 55° phase margin. What size resistor should be used to provide lead compensation with $\omega_z = 1.7\omega_t$? What is the final phase margin? What is the final unity-gain frequency and slew rate?

6.36 For the folded-cascode amplifier shown in Fig. 6.20, assume the bias currents of Q_1 and Q_2 are K times greater than the bias currents in Q_5 and Q_6. Assume the total bias current in the opamp, I_{total}, is determined by a specified power dissipation. Derive an equation for the unity-gain frequency in terms of K and I_{total}. Then show that the unity-gain frequency is maximized by taking K large.

6.37 Define K as in Problem 6.36, but this time derive the dc gain terms of K and I_{total} and show that it is maximized by taking K large.

6.38 Modify all of the device sizes of the opamp designed in Example 6.9 and compensated in Example 6.10 so that the gate lengths are reduced to $L = 0.3$ μm , while all values of V_{eff} and all drain currents are kept constant. What is the resulting effect on the opamp's dc gain, unity-gain frequency, and slew rate? Estimate the effect on the opamp's second equivalent pole frequency.

6.11.5 Section 6.5: Current Mirror Opamp

6.39 Derive an equation for the ratio of the unity-gain frequency of the folded-cascode amplifier of Fig. 6.20 to the unity-gain frequency of the current-mirror opamp of Fig. 6.22 in terms of K and I_{total} assuming both amplifiers have the same size input transistors, total power dissipation, and load capacitances. For the folded-cascode opamp, define K to be I_{D2}/I_{D7}. What is this ratio for K equal to 1, 2, and 4?

6.40 A current-mirror opamp is to be used in the feedback configuration of Fig. 6.23 with $C_1 = 4$ pF, $C_2 = 2$ pF, and C_p, C_c, C_{load} all negligible. Assume the input differential pair transistors are to be biased with $V_{eff,1} = 150$ mV, and $K = 3$. Estimate the total current consumption required to obtain a closed-loop 3-dB bandwidth of 60 MHz. How does this result compare with that obtained for the folded cascode opamp in Problem 6.31? What is the resulting slew rate?

6.41 Repeat Problem 6.40, but this time assume the input differential pair devices are biased in subthreshold so that $g_{m1} \approx qI_D/nkT$.

Table 6.3 The transistor sizes (in μm) for the opamp considered in Problem 6.44.

Q_1	90/0.2	Q_7	20/0.2	Q_{13}	4/0.2
Q_2	90/0.2	Q_8	60/0.2	Q_{14}	12/0.2
Q_3	20/0.2	Q_9	20/0.2		
Q_4	20/0.2	Q_{10}	60/0.2		
Q_5	20/0.2	Q_{11}	4/0.2		
Q_6	20/0.2	Q_{12}	12/0.2		

6.42 For the enhanced current-mirror circuit shown in Fig. P6.26, replace the amplifier A with the current-mirror opamp. Derive an expression for the output resistance, R_{out} in terms of the constituent transistors' small signal parameters (g_m, r_{ds}, etc.).

6.43 Resize transistors $Q_{8,\,10,\,12,\,14}$ in the current-mirror opamp of Example 6.11 to make $K = 4$. What is the resulting impact on dc gain, unity-gain frequency, and slew rate?

6.44 Consider the current mirror opamp of Fig. 6.22 with the device sizes listed in Table 6.3 and $I_b = 320\ \mu A$. What is the maximum value of C_L for which a unity-gain frequency of 80 MHz is provided? What is the resulting slew-rate?

6.45 It is desired to reduce the power consumption of the opamp in Problem 6.44 by 20% while maintaining the same slew-rate performance. Keeping all device lengths $L = 0.2\ \mu m$ and all V_{eff} the same, find new values for the device widths and I_b. For this new design, what is the maximum value of C_L for which a unity-gain frequency of 80 MHz is provided?

6.11.6 Section 6.6: Linear Settling Time Revisited

6.46 Consider the current-mirror opamp described in Example 6.11, in the feedback configuration shown in Fig. 6.23, with $C_1 = 1$ pF and $C_2 = C_C = C_{load} = 5$ pF. What is the linear settling time for 1 percent accuracy?

6.47 For the case described in Example 6.12, and assuming a current-mirror opamp described in Example 6.11, also assume the input voltage is a 1-V step change. Would the opamp slew rate limit? If so, how close must the output voltage come to its final value before its output voltage rate of change is less than the slew rate? How long would this take? How much longer would be required for linear settling to 1 percent of the total voltage change at the output?

6.48 Consider the circuit shown in Fig. P6.48, where a high-output-impedance opamp is used. Capacitor C_p is the parasitic capacitance at the opamp input, while C_o is the output capacitance. Assuming linear settling, show that the optimal value for minimizing the time constant of this circuit is given by

$$C_{1,opt} = \sqrt{MC_oC_p}$$

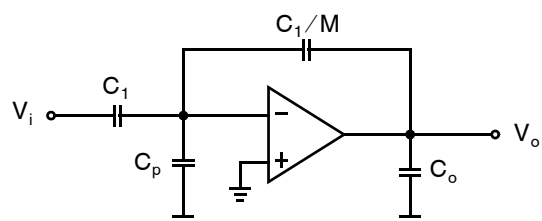


Fig. P6.48

6.49 Based on the results of Problem 6.48, derive τ_{min} for $M = 1$, $C_o = 1$ pF, and $C_p = 0.05$ pF. Sketch τ for other values of C_1 between 0.1 pF and 1 pF.

6.11.7 Section 6.7: Fully Differential Opamps

6.50 What are the single-ended slew rates of the fully differential folded-cascode and current-mirror opamps, assuming $K = 2$ and the current densities are the same as those of Example 6.9 and Example 6.11, respectively? Calculate the slew rates in both the positive and the negative directions. Assume the load capacitances are 10 pF between each output and ground. For the folded-cascode design, assume the clamp transistors Q_{11} and Q_{12} of Fig. 6.28 have not been included. Also, assume the current sources biasing the output stages do not change during transients.

6.51 Derive the slew rate of the fully differential opamp of Fig. 6.30 in terms of C_L, K, and I_{bias}.

6.52 Derive an equation for the unity-gain frequency of the fully differential opamp shown in Fig. 6.31 in terms of K and I_{total}.

6.53 Derive expressions for the unity-gain frequency and slew rate of the fully differential opamp shown in Fig. 6.33 in terms device parameters and sizes, component values, and the bias currents I_1, I_2, and I_3.

6.11.8 Section 6.8: Common-Mode Feedback Circuits

6.54 Assume the CMFB circuit of Fig. 6.34 has a 1.8-V power supply voltage, and that the current sources require 0.3 V across them in order to have all transistors remain in the active regions. Ignoring the body effect, what V_{eff} bias voltage should be used for the p-channel transistors to maximize signal swing? What is the maximum single-ended signal swing before the gain of the common-mode feedback circuitry goes to zero and why?

6.55 Repeat Problem 6.54, but do not ignore the body effect. Assume the n-wells of the p-channel transistors are connected to V_{DD}.

6.56 For the circuit of Fig. 6.37, find an expression for the gain from V_{cntrl} to the output common-mode voltage, $V_{out,cm}$, in terms of the transistor small-signal parameters. If the common-mode sense and different amplifier circuits of Fig. 6.36 are used in Fig. 6.37, find an expression for the loop gain.

6.57 A simple common-mode feedback circuit for a folded-cascode opamp is presented in Fig. P6.57. If the bias currents of all transistors are $I_D = 100\ \mu A$ and device lengths are $L = 0.25\ \mu m$, choose the device widths W_{7-10} identically so that the output common-mode level is 0.8 V. Next, select V_{B3} to maximize the available output signal swing.

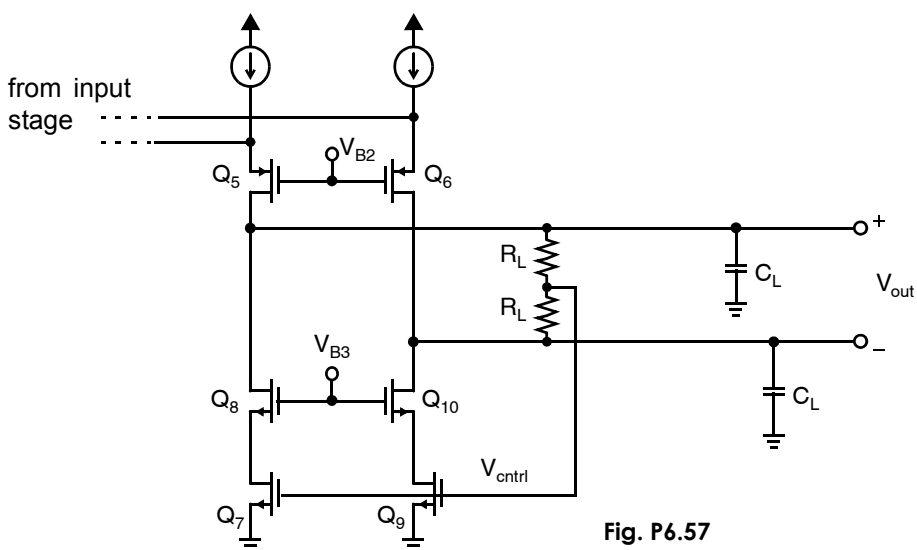


Fig. P6.57

CHAPTER

7 Biasing, References, and Regulators

Although often ignored during the course of first-pass analog design, a critical factor in determining a circuit's overall performance is the quality of the dc voltage and current sources. This chapter covers the design of circuits used to establish those dc voltages or currents. They are, themselves, sophisticated analog circuits usually employing feedback.

7.1 ANALOG INTEGRATED CIRCUIT BIASING

In an analog integrated circuit, many subcircuits work together to generate all of the various dc voltages and currents. These include bias circuits, reference circuits, and regulators. A *bias* circuit generates the dc voltages required to keep transistors near some desired operating point; of course, as transistor parameters change, either from chip to chip or with changes in temperature, so must the bias voltages. A *reference* circuit generates a voltage and/or current of a known fixed absolute value (for example, one volt). Finally, a *regulator* circuit improves the quality of a dc voltage or current, usually decreasing the noise. Fig. 7.1 shows how these circuits may work together to support the analog circuits on a large mixed analog–digital chip.

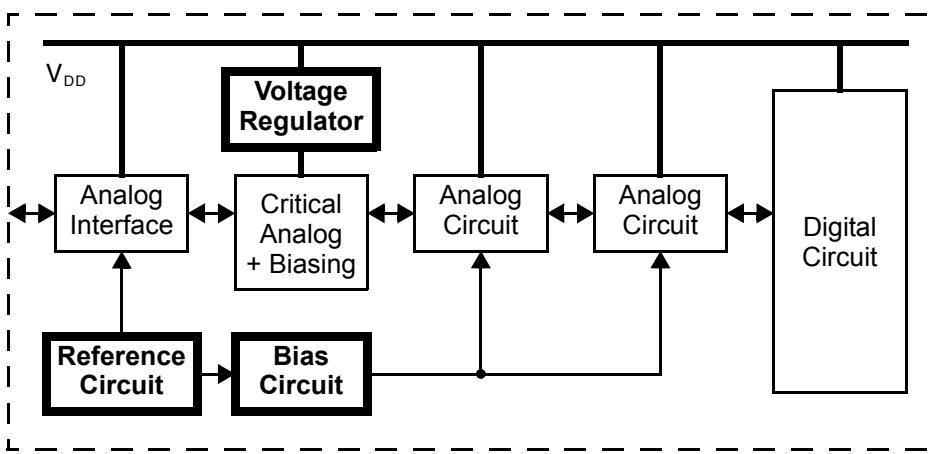


Fig. 7.1 A large mixed analog–digital integrated circuit emphasizing the role of biasing, references, and regulators.

7.1.1 Bias Circuits

Generally, the objective of a bias circuit is to ensure that the dc operating points of the transistors in an analog circuit remain within a desired range. Unlike a reference circuit, this means that the dc outputs of a bias circuit may adjust to process and temperature variations. Several approaches may be taken towards biasing. One is to ensure that circuits are biased to permit constant voltage swings; another is to ensure that constant currents are maintained; yet another is to try to ensure constant gain is maintained. These are illustrated in the following example.

> ***Key Point:*** *A bias circuit generates the dc voltages required to keep transistors near some desired operating point; as transistor parameters change, either from chip to chip or with changes in temperature, so must the bias voltages.*

EXAMPLE 7.1

Consider the simple NMOS differential pair in Fig. 7.2 biased, nominally, with $V_{eff,1} = 200$ mV and $V_{tn} = 450$ mV. With a 10% decrease in both $\mu_n C_{ox}$ and R, and a 10% increase in V_{tn}, how must V_b change to ensure: a) a constant drain current in the matched differential pair devices $Q_{2,3}$; b) a constant voltage drop across the resistors R; c) a constant gain.

Solution

Assume all MOSFETs are in active mode and obeying a simple square-law voltage–current relationship. Before the 10% changes, $V_b = V_{tn} + V_{eff,1} = 650$ mV.

a) The drain currents of $Q_{2,3}$ are one-half the drain current in Q_1,

$$I_{D1} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_b - V_{tn})^2.$$

Hence, to maintain a constant drain current the gate voltage V_b will have to increase 45 mV along with V_{tn}. In addition, the effective gate-source voltage $V_{eff,1} = V_b - V_{tn}$ must change in inverse proportion to changes in $\sqrt{\mu_n C_{ox}}$. However, R has only secondary effects on drain current, so little if any change is required in V_b in response to small changes in R. Therefore,

$$V_b = V_{tn} + V_{eff,1} = 1.1(450\text{ mV}) + (200\text{ mV})/(\sqrt{0.9}) = 706\text{ mV}.$$

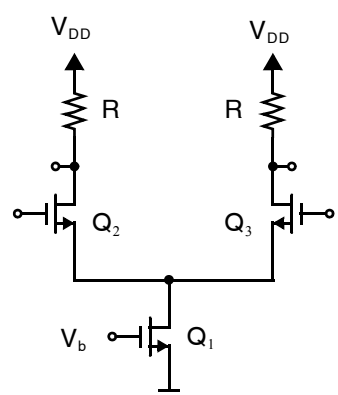


Fig. 7.2 NMOS differential pair biased by a simple current source.

b) The voltage drop across R is given by $I_{D1}R/2$. In addition to the dependencies described in part (a) above, keeping the voltage drop constant will require the drain current, and hence $(V_b - V_{tn})$, to change in inverse proportion to changes in R. Therefore,

$$V_b = 1.1(450\text{ mV}) + (200\text{ mV})/(0.9\sqrt{0.9}) = 729\text{ mV}.$$

c) The small-signal gain of the differential pair is $g_{m2}R$ where it may be shown that

$$g_{m2} \propto \mu_n C_{ox}(V_b - V_{tn})$$

Hence, as in parts (a) and (b), keeping the gain constant demands that V_b change in direct proportion to any changes in V_{tn}. Moreover, $V_{eff,1} = V_b - V_{tn}$ must change in inverse proportion to changes in $\mu_n C_{ox}$ and R. Therefore,

$$V_b = 1.1(450\text{ mV}) + (200\text{ mV})/(0.9 \cdot 0.9) = 742\text{ mV}$$

Example 7.1 illustrates the challenge of designing a good bias circuit which must somehow monitor multiple device parameters and automatically adjust multiple analog voltages to achieve a desired effect.

Since bias circuits only support other analog circuits by providing their dc voltages, their area and power consumption represents an overhead cost that a designer will want to minimize. Hence, one bias circuit is usually shared by several subcircuits. An important practical consideration is, therefore, how to distribute the bias circuit outputs without subjecting them to noise and inaccuracies due to component mismatches across the chip.

Most (or all) of a bias circuit's outputs will be the gate (or base) voltages used to bias current sources. When distributing these bias voltages, any devices that are required to operate as a current mirror should be placed within the same subcircuit in close physical proximity to one another to ensure good matching between those devices.

For example, consider the common scenario depicted in Fig. 7.3: a reference current, I_b, is available at one location on an integrated circuit and must be replicated at the drain of NMOS device Q_2 located at a great distance across the chip. Two methods for generating the gate voltage of Q_2 are shown. In Fig. 7.3(*a*), a simple NMOS current mirror is used and the mirror devices are separated by a long wire. In this scheme, the long wire carries zero current and it is hence referred to as "voltage-mode" bias distribution. Inevitable differences between the threshold voltages of Q_1 and Q_2, and even the ground potentials at their source and body terminals lead to considerable uncertainty in the resulting current I_c.

In Fig. 7.3(*b*), PMOS current mirror $Q_{3/4}$ is introduced to generate an intermediate current I_d that is transmitted across the chip. Unlike Fig. 7.3(*a*), here the long wire carries a non-zero current making this a "current-mode" approach to bias distribution. It is preferred since both pairs of current mirror devices are co-located, so that their currents may be well-matched. Naturally, the simple current mirrors in Fig. 7.3 may be replaced by high output impedance variants to improve their accuracy, but the basic advantage of current-mode bias distribution remains. Current-mode distribution may consume more power and area than voltage-mode distribution due to the additional devices and bias current required. However, these minor disadvantages are ameliorated by recognizing that Q_1 and Q_3 in Fig. 7.3(*b*) can be made quite small and the current I_d reduced so that the added power and area are minimal.[1]

Also note that Fig. 7.3(*b*) includes decoupling capacitors at the gates of all transistors. These are important to filter noise at those nodes and may be implemented as MOS capacitors. The decoupling capacitor is connected to

---

1. A practical lower limit of 10's of μA is usually established for I_d since smaller currents can result in too much noise on the resulting bias voltages.

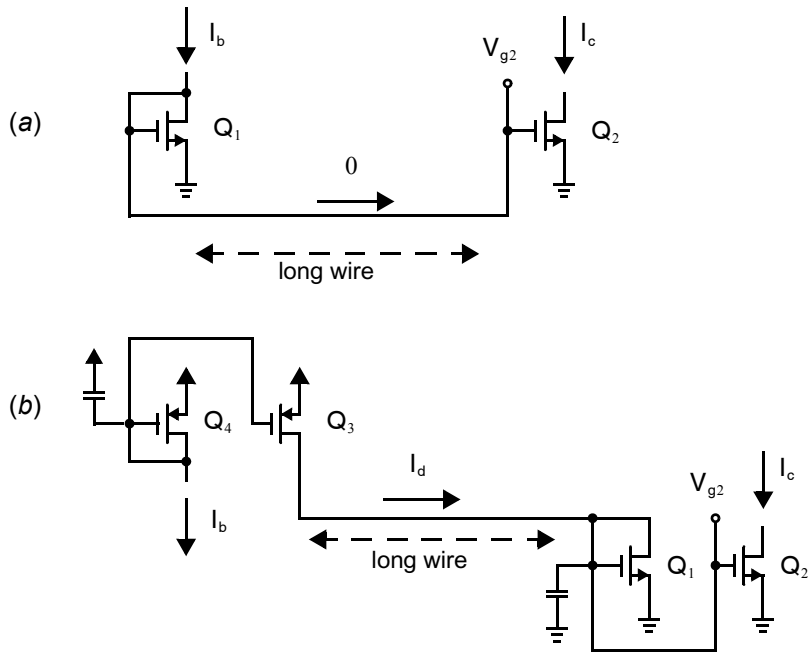


Fig. 7.3 Two approaches to bias distribution: (*a*) voltage-mode (not recommended); (*b*) current-mode.

the positive supply, not ground, because its purpose is to maintain a constant gate-source voltage on Q_3 and Q_4. Hence, it is desirable for that gate voltage to track any noise on the positive supply, thereby maintaining a constant bias current I_d.

7.1.2 Reference Circuits

Key Point: *A reference circuit generates a voltage and/or current of a known fixed absolute value.*

Known absolute values of voltage or current are most useful at the interface between integrated circuits, or between an integrated circuit and some other discrete component. For example, two integrated circuits may be required to interface with a signal swing of one volt. A reference voltage or current may sometimes be derived from the supply voltage, but the supply is not always controlled with sufficient accuracy in which case a reference voltage or current must be produced by an integrated reference circuit. Unfortunately, although dimensionless quantities may be accurately controlled on integrated circuits (e.g., ratios of device sizes), there are very few dimensioned quantities that do not vary significantly from one integrated circuit to another, or with variations in temperature.

EXAMPLE 7.2

Transistor Q_1 in Fig. 7.4 has an aspect ratio of $W/L = 10$, $\mu_n C_{ox} = 190\ \mu A/V^2$ and $V_{tn} = 0.45$ V. Resistor $R = 20k\Omega$ and $V_{DD} = 1.5V$. How much will V_o change if the value of $\mu_n C_{ox}$ decreases by 30%, R decreases by 10%, V_{tn} increases by 30mV, and V_{DD} increases by 100 mV?

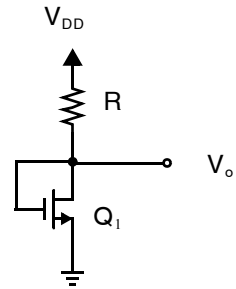


Fig. 7.4 A simple circuit to establish a dc voltage at V_o.

Solution

Assuming a square-law model for Q_1 and neglecting its finite output impedance to get an approximate answer, V_o may be found by solving the following equation.

$$\frac{V_{DD} - V_o}{R} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_o - V_{tn})^2 \tag{7.1}$$

Equation (7.1) is a quadratic in V_o that has only one valid solution with $V_o > V_{tn}$.

$$V_o = V_{tn} + \frac{1}{\mu_n C_{ox}(W/L)R}[\sqrt{2(V_{DD} - V_{tn})\mu_n C_{ox}(W/L)R + 1} - 1] \tag{7.2}$$

Substituting the nominal values given above into (7.2) yields $V_o = 660$ mV. Making the prescribed changes to the circuit parameters results in $V_o = 747$ mV. This represents an increase of 13% under variations that are quite typical.

The most widely used quantity for generating a reference of high accuracy in integrated circuits (on the order of 1% or better) has proven to be the bandgap of silicon. It is a constant whose value does not change with variations in temperature, dopant concentrations, or device dimensions. Circuits that produce a voltage proportional to the bandgap of silicon are described in Section 7.3.

7.1.3 Regulator Circuits

Key Point: *A regulator circuit improves the quality of a dc voltage or current, usually decreasing the noise.*

A regulator's main purpose is to produce a voltage which has low noise and from which some current may be drawn. They are common, for example, when a critical analog circuit must operate from the same supply voltage as other circuits. As shown in Fig. 7.5, the other circuits introduce significant noise onto the common supply, but a regulator can maintain a quiet supply for the critical circuitry. Digital circuits in particular are major sources of power supply noise, so regulators are common in mixed analog–digital systems.

The general approach is to use a feedback amplifier operating under the noisy supply to generate a quiet dc voltage that supplies the critical circuit. Hence, the regulated voltage is generally lower than the regulator's supply voltage. Important specifications are the immunity of the regulator's output to variations in supply voltage and in the load current.

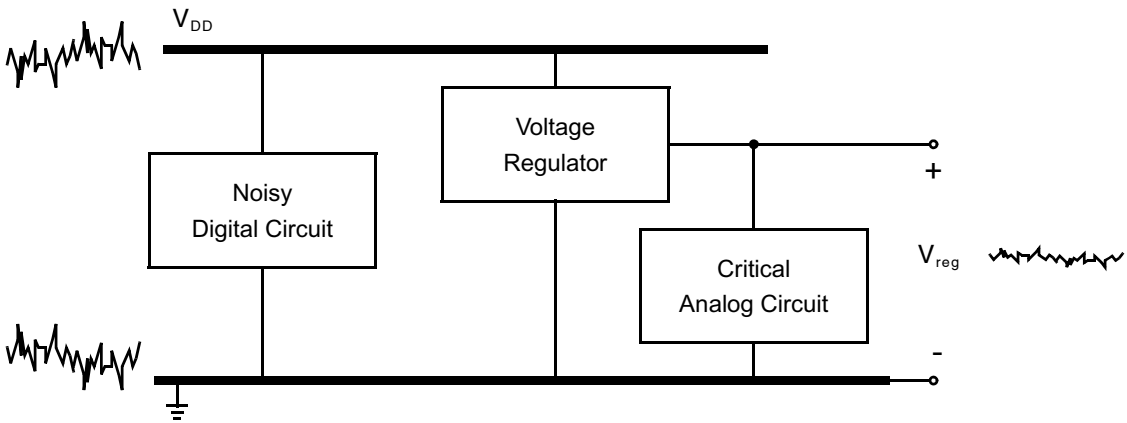


Fig. 7.5 A regulator maintains a quiet supply voltage for a critical analog circuit within a larger system having a noisy supply voltage.

7.2 ESTABLISHING CONSTANT TRANSCONDUCTANCE

7.2.1 Basic Constant-Transconductance Circuit

We have seen that transistor transconductances are perhaps the most important parameters in analog amplifiers that must be stabilized. This stabilization can be achieved by using a circuit approach first proposed in [Steininger, 1990] in which transistor transconductances are matched to the conductance of a resistor. As a result, to a first-order effect, the transistor transconductances are independent of power-supply voltage as well as process and temperature variations.

The bias circuit is shown in Fig. 7.6. First it is assumed that $(W/L)_{10} = (W/L)_{11}$. This equality results in both sides of the circuit having the same current due to the current-mirror pair Q_{10}, Q_{11}. As a result, we also must have $I_{D15} = I_{D13}$. Now, around the loop consisting of Q_{13}, Q_{15}, and R_B, we have

$$V_{GS13} = V_{GS15} + I_{D15}R_B \tag{7.3}$$

and recalling that $V_{effi} = V_{GSi} - V_t$, we can subtract the threshold voltage, V_t, from both sides, resulting in

$$V_{eff13} = V_{eff15} + I_{D15}R_B \tag{7.4}$$

This equation can also be written as

$$\sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}} = \sqrt{\frac{2I_{D15}}{\mu_n C_{ox}(W/L)_{15}}} + I_{D15}R_B \tag{7.5}$$

and since $I_{D13} = I_{D15}$, we can also write

$$\sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}} = \sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{15}}} + I_{D13}R_B \tag{7.6}$$

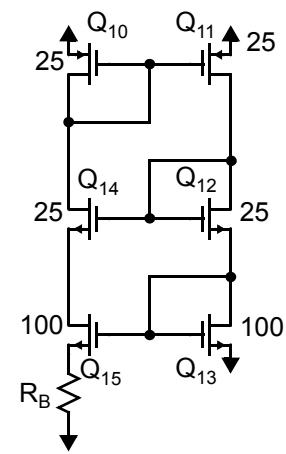


Fig. 7.6 A bias circuit that gives very predictable and stable transistor transconductances, especially for n-channel devices.

Rearranging, we obtain

$$\frac{2}{\sqrt{2\mu_n C_{ox}(W/L)_{13} I_{D13}}}\left[1 - \sqrt{\frac{W/L_{13}}{W/L_{15}}}\right] = R_B \tag{7.7}$$

and recalling that $g_{m13} = \sqrt{2\mu_n C_{ox}(W/L)_{13} I_{D13}}$ results in the important relationship

$$g_{m13} = \frac{2\left[1 - \sqrt{\dfrac{(W/L)_{13}}{(W/L)_{15}}}\right]}{R_B} \tag{7.8}$$

Key Point: *A constant-transconductance bias circuit produces a small-signal transconductance that is a fixed fraction of a resistor value, independent of process, temperature, or supply variations.*

Thus, the transconductance of Q_{13} is determined by R_B and geometric ratios only, independent of power-supply voltages, process parameters, temperature, or any other parameters with large variability. For the special case of $(W/L)_{15} = 4(W/L)_{13}$, we have simply

$$g_{m13} = \frac{1}{R_B} \tag{7.9}$$

Not only is g_{m13} stabilized, but all other transconductances are also stabilized since all transistor currents are derived from the same biasing network, and, therefore, the ratios of the currents are mainly dependent on geometry. For example, for all n-channel transistors,

$$g_{mi} = \sqrt{\frac{(W/L)_i I_{Di}}{(W/L)_{13} I_{D13}}} \times g_{m13} \tag{7.10}$$

and for all p-channel transistors

$$g_{mi} = \sqrt{\frac{\mu_p}{\mu_n}\frac{(W/L)_i I_{Di}}{(W/L)_{13} I_{D13}}} \times g_{m13} \tag{7.11}$$

Key Point: *The constant-transconductance bias circuit ensures a constant transconductance in all transistors of the same type as Q_{13}. However, the transconductance of transistors of the opposite type will vary.*

Since g_{m13} is an n-channel device, the additional term $\sqrt{\mu_p/\mu_n}$ appears in (7.11) modifying the transconductance of all p-channel devices. This, unfortunately, exhibits large variations from chip-to-chip and small variations with temperature.

The circuit in Fig. 7.6 forms a positive feedback loop, and stability is a significant concern. Transistor Q_{11} acts as a common-source amplifier for any small signals appearing at its gate, having diode-connected Q_{13} as a load. The resulting signal then appears at the gate of Q_{15}, another common-source stage with R_B serving as a degeneration resistor. As long as R_B is large enough, this second stage will have a gain much less than unity, ensuring stability of the loop. However, at high frequencies any parasitic capacitance across R_B will reduce its impedance and tend to make the circuit unstable [Nicolson, 2004]. If R_B is to be a precision off-chip resistor, the parasitic capacitances will be large including the chip pad and board. Therefore, it is usually necessary to implement at least part of the resistance R_B on-chip. Unfortunately, on-chip resistances are relatively poorly controlled.

The preceding analysis has ignored many second-order effects such as the transistor output impedance and the body effect. The body effect will modify the equation slightly, but the relationship will still depend primarily on geometry alone. The major limitation is due to the transistor output impedance. Both effects can be mitigated using the modified circuit shown in Fig. 7.7. Here, the pair $Q_{14,15}$ is implemented with PMOS devices which in most CMOS processes permits connection to an independent body terminal. Connecting the source and body of Q_{15} together eliminates the body effect. Furthermore, an amplifier is used to maintain equal voltages at the transistor drain terminals, thus reducing the effect of finite transistor output impedance. Another advantage of the amplifier is that it reduces the impedance at the drain of Q_{15}, which will reduce the gain around the positive feedback loop Q_{12-15} and improve stability. The negative feedback loop comprising the amplifier and common-source transistors $Q_{12,13}$ may be stabilized by proper sizing of compensation capacitor C_C.

Unfortunately, both Fig. 7.6 and Fig. 7.7 have a second stable state where all the currents are zero. To guarantee this condition doesn't happen, it is necessary to add a *start-up circuit* that only affects the operation if all the currents are zero at start up.

A fundamental limitation of constant-transconductance biasing is that at high temperatures, the currents and effective gate-source voltages increase substantially to compensate for decreasing carrier mobility and keep the transconductances stable. This limits signal swings which is particularly problematic under low supply voltages. Since the carrier mobility is proportional to $T^{-3/2}$, this corresponds to a 27-percent reduction from room temperature (300 K) to 100 °C (373 K). Thus, the effective gate-source voltages increase by 27 percent to keep the transistor transconductance unchanged since

$$g_{mi} = \mu_i C_{ox}\left(\frac{W}{L}\right)_i V_{eff\text{-}i} \tag{7.12}$$

As long as the effective gate-source voltages have not initially been designed to be too large, this limitation is tolerable in most applications.[2] A typical value for effective gate-source voltages might be 0.2 V to 0.25 V at room temperature.

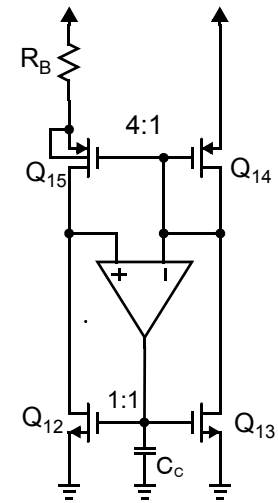


Fig. 7.7 A modified bias circuit for giving predictable and stable transistor transconductances.

7.2.2 Improved Constant-Transconductance Circuits

It is possible to incorporate wide-swing current mirrors into the constant-transconductance bias circuit described above. This modification greatly minimizes most of the detrimental second-order imperfections caused by the finite-output impedance of the transistors, without greatly restricting signal swings. The complete circuit is shown in Fig. 7.8 [McLaren, 2001]. This circuit is a modification of the circuit described in Fig. 7.6, and has both wide-swing current mirrors and a start-up circuit.

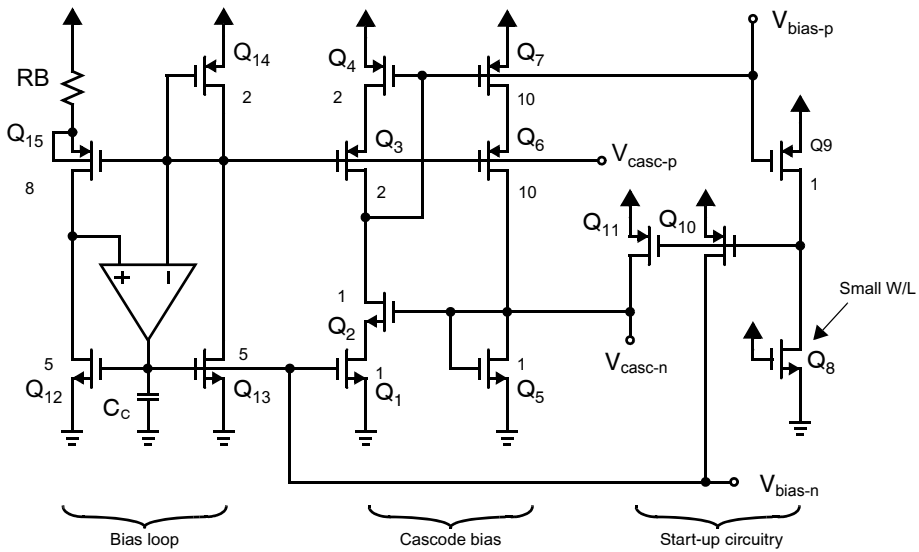
Key Point: *Many bias circuits require auxiliary start-up circuits to prevent them from resting in a zero-current state indefinitely.*

A constant-transconductance bias is provided by transistors Q_{12-15}, so that $g_{m14} = 1/R_B$ as in Fig. 7.7. The resulting current is then mirrored to Q_1 and used to the generate the PMOS gate bias voltage $V_{bias\text{-}p}$. Notice that the current density $I_D/(W/L)$ of Q_{13} is five times greater than that of Q_4. Hence, the gate voltage of Q_{13} is suitable for use as the PMOS cascode bias $V_{casc\text{-}p}$. Similarly, $Q_{3,4,6,7}$ form a wide-swing PMOS cascode current mirror that directs current into Q_5 with five times greater current density than in Q_1, so that the gate voltage V_{G5} is a suitable NMOS cascode bias $V_{casc\text{-}n}$.

An example of a start-up circuit is also shown on the right side of Fig. 7.8. In the event that all currents in the bias loop are zero, Q_9 will be off. Since Q_8 operates as a high-impedance load that is always on, the gates of $Q_{10,11}$ will be pulled low. These transistors then will inject currents into the bias loop, which will start up the circuit. Once the loop starts up, Q_9 will turn on, sourcing all of the current for Q_8, pulling the gates of $Q_{10,11}$ high, and thereby turning them off so they no longer affect the bias loop. This circuit is only one example of a start-up loop, and there are many other variations. For example, sometimes the n-channel transistor, Q_8, is replaced by an actual resistor (perhaps realized using a well resistor).

It is of interest to note that the bias circuit shown consists of four different loops—the main loop with positive feedback, the start-up loop that eventually gets disabled, and the two loops used for establishing the bias voltages for the cascode transistors. These latter two loops also constitute positive feedback but with very little gain. Finally, note that the opamp included in the constant-g_m bias loop may itself be biased by the loop.

2. Also, if on-chip resistors, such as well or diffusion resistors, are used, this effect will be somewhat offset by their positive temperature-coefficient dependency.



SPICE! *Simulate this circuit with the provided netlist.*

Fig. 7.8 A constant-transconductance bias circuit having wide-swing cascode current mirrors. Shown next to each device are reasonable choices for the device widths normalized to the width of Q_1 assuming equal gate lengths.

7.3 ESTABLISHING CONSTANT VOLTAGES AND CURRENTS

An important analog building block, especially in data acquisition systems, is a voltage reference. Ideally, this block will supply a fixed dc voltage of known amplitude that does not change with temperature. This can be combined with an accurate resistance to provide a stable dc current, if needed. There have been a number of approaches that have been taken to realize voltage references in integrated circuits. These include,

1. Making use of a zener diode that breaks down at a known voltage when reverse biased.
2. Making use of the difference in the threshold voltage between an enhancement transistor and a depletion transistor.
3. Cancelling the negative temperature dependence of a pn junction with a positive temperature dependence from a PTAT (proportional-to-absolute-temperature) circuit.

The first approach is no longer popular because the breakdown voltage of a zener diode is typically larger than the power supplies of modern integrated circuits. The second approach cannot be used when depletion transistors are not available, as is often the case. In addition, although it can be used to make quite stable references, the actual value of the reference is difficult to determine accurately because of the process sensitivity of the difference between the threshold voltage of an enhancement device and a depletion device. For these reasons, the first two approaches are not covered here. Rather, the last approach, which is currently the most popular for both bipolar and CMOS technologies, will be discussed. Voltage references based on the last approach are commonly called "bandgap" voltage references for reasons that will become apparent shortly.

7.3.1 Bandgap Voltage Reference Basics

As just mentioned, a bandgap voltage reference is based on subtracting the voltage of a forward-biased diode (or base-emitter junction) having a negative temperature coefficient from a voltage proportional to absolute temperature

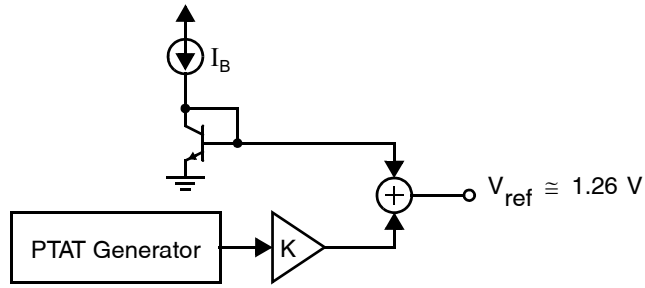


Fig. 7.9 A simplified circuit of a bandgap voltage reference.

(PTAT). As we shall see, this PTAT voltage is realized by amplifying the voltage difference of two forward-biased base-emitter (or diode) junctions. A bandgap voltage reference system is shown symbolically in Fig. 7.9.

A forward-biased base-emitter junction of a bipolar transistor[3] has an I-V relationship given by

$$I_C = I_S e^{qV_{BE}/kT} \tag{7.13}$$

where I_S is the transistor scale current and, although not shown, has a strong dependence on temperature.

Writing the base-emitter voltage as a function of collector current and temperature, it can be shown that [Brugler, 1967; Tsividis, 1980]

$$V_{BE} = V_{G0}\left(1 - \frac{T}{T_0}\right) + V_{BE0}\frac{T}{T_0} + \frac{mkT}{q}\ln\left(\frac{T_0}{T}\right) + \frac{kT}{q}\ln\left(\frac{J_C}{J_{C0}}\right) \tag{7.14}$$

Here, V_{G0} is the bandgap voltage of silicon extrapolated to 0 K (approximately 1.206 V), k is Boltzmann's constant, and m is a temperature constant approximately equal to 2.3. Also, J_C and T are the collector current density and temperature, respectively, while the subscript 0 designates an appropriate quantity at a reference temperature, T_0. Specifically, J_{C0} is the collector current density at the reference temperature, T_0, whereas J_C is the collector current density at the true temperature, T. Also, V_{BE0} is the junction voltage at the reference temperature, T_0, whereas V_{BE} is the base-emitter junction voltage at the true temperature, T. Note that the junction current is related to the junction current density according to the relationship

$$I_C = A_E J_C \tag{7.15}$$

where A_E is the effective area of the base-emitter junction.

For I_C constant, V_{BE} will have approximately a -2 mV/°K temperature dependence around room temperature. This negative temperature dependence is cancelled by a PTAT temperature dependence of the amplified difference of two base-emitter junctions biased at fixed but different current densities. Using (7.14), it is seen that if there are two base-emitter junctions biased at currents J_2 and J_1, then the difference in their junction voltages is given by

$$\Delta V_{BE} = V_2 - V_1 = \frac{kT}{q}\ln\left(\frac{J_2}{J_1}\right) \tag{7.16}$$

Key Point: *An integrated voltage reference is made by adding the forward bias voltage of a pn junction to the difference in the forward voltages of two pn junctions biased at different current densities. Before adding them, these voltages are scaled so that their respective positive and negative temperature coefficients cancel precisely.*

Thus, the difference in the junction voltages is proportional to absolute temperature. This proportionality is quite accurate and holds even when the collector currents are temperature dependent, as long as their ratio remains fixed.

3. It is here assumed the reader is familiar with the basics of bipolar transistors. If needed, the reader may refer to Section 8.1.1 for this background.

EXAMPLE 7.3

Assume two transistors are biased at a current-density ratio of 10:1 at $T = 300$ K. What is the difference in their base-emitter voltages and what is its temperature dependence?

Solution

Using (7.16), we have

$$\Delta V_{BE} = \frac{kT}{q} \ln\left(\frac{J_2}{J_1}\right) = \frac{1.38 \times 10^{-23}(300)}{1.602 \times 10^{-19}} \ln(10) = 59.5 \text{ mV} \tag{7.17}$$

Since this voltage is proportional to absolute temperature, after a 1 K temperature increase, the voltage difference will be

$$\Delta V_{BE} = 59.5 \text{ mV} \frac{301}{300} = 59.7 \text{ mV} \tag{7.18}$$

Thus, the voltage dependence is 59.5 mV/300 K or 0.198 mV/K. Since the temperature dependence of a single V_{BE} is −2 mV/K, if it is desired to cancel the temperature dependence of a single V_{BE}, then ΔV_{BE} should be amplified by about a factor of 10, as explained next.

It will be seen shortly that when realizing a bandgap voltage reference, although the output voltage is temperature independent, the junction currents turn out to be proportional to absolute temperature (assuming the resistors used are temperature independent). Thus, to simplify derivations, we will first assume the junction currents are proportional to absolute temperature. Later, it will be verified that this proportionality relationship is true when circuit realizations are described. We therefore first assume

$$\frac{J_i}{J_{i0}} = \frac{T}{T_0} \tag{7.19}$$

where J_i is the current density of the collector current of the i^{th} transistor, whereas J_{i0} is the same current density at the reference temperature.

Now, assume that the difference between two base-emitter voltages is multiplied by a factor of K and added to the base-emitter voltage of the junction with the larger current density. Using (7.16) and (7.19) along with (7.14), we have

$$\begin{aligned} V_{ref} &= V_{BE2} + K\Delta V_{BE} \\ &= V_{G0} + \frac{T}{T_0}(V_{BE0\text{-}2} - V_{G0}) + (m-1)\frac{kT}{q}\ln\left(\frac{T_0}{T}\right) + K\frac{kT}{q}\ln\left(\frac{J_2}{J_1}\right) \end{aligned} \tag{7.20}$$

Equation (7.20) is the fundamental equation giving the relationship between the output voltage of a bandgap voltage reference and temperature. Here, $V_{BE0\text{-}2}$ is the base-emitter junction voltage of the second transistor at temperature T_0. If we want zero temperature dependence at a particular temperature, we can differentiate (7.20) with respect to temperature and set the derivative to zero at the desired reference temperature. From (7.20), we have

$$\frac{\partial V_{ref}}{\partial T} = \frac{1}{T_0}(V_{BE0\text{-}2} - V_{G0}) + K\frac{k}{q}\ln\left(\frac{J_2}{J_1}\right) + (m-1)\frac{k}{q}\left[\ln\left(\frac{T_0}{T}\right) - 1\right] \tag{7.21}$$

Setting (7.21) equal to zero at $T = T_0$, we see that for zero temperature dependence at the reference temperature, we need

$$V_{BE0\text{-}2} + K\frac{kT_0}{q}\ln\left(\frac{J_2}{J_1}\right) = V_{G0} + (m-1)\frac{kT_0}{q} \tag{7.22}$$

The left side of (7.22) is the output voltage V_{ref} at $T = T_0$ from (7.20). Thus for zero temperature dependence at $T = T_0$, we need

$$V_{ref\text{-}0} = V_{G0} + (m-1)\frac{kT_0}{q} \tag{7.23}$$

For the special case of $T_0 = 300\ ^\circ\text{K}$ and $m = 2.3$, (7.23) implies that

$$V_{ref\text{-}0} = 1.24\ \text{V} \tag{7.24}$$

for zero temperature dependence. Notice that this value is independent of the current densities chosen. Thus, if a larger current density is chosen, then K must be taken appropriately smaller to achieve the correct reference output voltage. In precision integrated voltage references, this correct output voltage is achieved by trimming at the time the wafer is being tested. From (7.22), the required value for K is

$$K = \frac{V_{G0} + (m-1)\dfrac{kT_0}{q} - V_{BE0\text{-}2}}{\dfrac{kT_0}{q}\ln\left(\dfrac{J_2}{J_1}\right)} = \frac{1.24 - V_{BE0\text{-}2}}{0.0258\ \ln\left(\dfrac{J_2}{J_1}\right)} \tag{7.25}$$

at $300\ ^\circ\text{K}$.

The reason for the name of the bandgap voltage should now be apparent. Specifically, for zero temperature dependence, the output of a bandgap voltage reference is given by the bandgap voltage plus a small correction term to account for second-order effects.

The output voltage of the reference for temperatures different from the reference is found after back-substituting (7.22) and (7.23) into (7.20). After some manipulations, the result is

$$V_{ref} = V_{G0} + (m-1)\frac{kT}{q}\left[1 + \ln\left(\frac{T_0}{T}\right)\right] \tag{7.26}$$

and

$$\frac{\partial V_{ref}}{\partial T} = (m-1)\frac{k}{q}\ln\left(\frac{T_0}{T}\right) \tag{7.27}$$

These equations can be used to estimate the temperature dependence at temperatures different from the reference temperature. In the next section, a practical bipolar realization of a bandgap reference will be described.

EXAMPLE 7.4

Estimate the temperature dependence at $0\,^\circ\text{C}$ for a bandgap voltage reference that was designed to have zero temperature dependence at $20\,^\circ\text{C}$. Present the result as ppm/K.

Solution

Recalling that 0°K corresponds to -273°C, we can write $T_0 = 293$ K and $T = 273$ K. Substituting these values into (7.27), we have

$$\frac{\partial V_{ref}}{\partial T} = (2.3 - 1)\frac{1.38 \times 10^{-23}}{1.6 \times 10^{-19}} \ln\left(\frac{293}{273}\right) = 8\ \mu V/K \tag{7.28}$$

For a reference voltage of 1.24 V, a dependency of $8\ \mu V/K$ results in

$$\frac{8\ \mu V/°K}{1.24\ V} = 6.5 \times 10^{-6}\ \text{parts}/°K = 6.5\ \text{ppm}/K \tag{7.29}$$

where ppm represents parts per million. It should be mentioned here that practical effects result in voltage references with typically 4 to 10 times larger values than this small amount of temperature dependency. It should also be noted that the ideal first-order temperature dependence of this bandgap voltage circuit is $0\ \text{ppm}/K$ at the reference temperature of 20 °C.

7.3.2 Circuits for Bandgap References

Bipolar Bandgap References

A voltage reference originally proposed in [Brokaw, 1974] has been the basis for many bipolar bandgap references. A simplified schematic of the circuit is shown in Fig. 7.10. The amplifier in the feedback loop keeps the collector voltages of Q_1 and Q_2 equal. Since $R_3 = R_4$, this guarantees that both transistors have the same collector currents and collector-emitter voltages. Also, notice that the emitter area of Q_1 has been taken eight times larger than the emitter area of Q_2. Therefore, Q_2 has eight times the current density of Q_1, resulting in

$$\frac{J_2}{J_1} = 8 \tag{7.30}$$

We have for the circuit

$$V_{ref} = V_{BE2} + V_{R1} \tag{7.31}$$

Also

$$\begin{aligned} V_{R1} &= I_{R1}R_1 \\ &= 2I_{R2}R_1 \end{aligned} \tag{7.32}$$

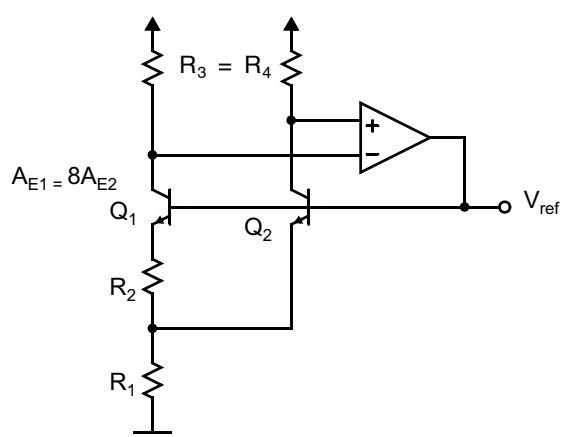


Fig. 7.10 A simplified schematic of a bipolar bandgap voltage reference.

But

$$I_{R2} = \frac{V_{R2}}{R_2} = \frac{V_{BE2} - V_{BE1}}{R_2} = \frac{\Delta V_{BE}}{R_2} \tag{7.33}$$

Substituting (7.32) and (7.33) into (7.31) gives

$$V_{ref} = V_{BE2} + \frac{2R_1}{R_2}\,\Delta V_{BE} \tag{7.34}$$

which is of the form desired to realize a bandgap reference. It is immediately recognizable that

$$K = \frac{2R_1}{R_2} \tag{7.35}$$

Assuming $V_{BE2\text{-}0} = 0.65$ V, from (7.25)

$$\frac{R_1}{R_2} = \frac{1}{2} \times \frac{1.24 - 0.65}{0.0258 \times \ln(8)} = 5.5 \tag{7.36}$$

In an integrated implementation, R_1 or R_2 would be trimmed while monitoring V_{ref} to force it equal to the desired reference voltage. Furthermore, the optimum value for this voltage might be determined empirically during the prototype phase of the design cycle.

Notice also, from (7.16) and (7.33), we have

$$I_{E1} = I_{E2} = I_{R2} = \frac{\Delta V_{BE}}{R_2} = \frac{\frac{kT}{q}\ln\left(\frac{J_2}{J_1}\right)}{R_2} \tag{7.37}$$

implying that all currents are proportional to absolute temperature (assuming that resistor R_2 is temperature independent). Thus, as assumed earlier, all currents are indeed PTAT. It is worth mentioning here that PTAT currents are often used to bias many bipolar circuits, as they result in transistor transconductances being independent of temperature. This constant transconductance has the desirable feature that circuit speed is relatively independent of temperature, but, unfortunately, has the undesirable feature that the circuit power dissipation goes up considerably at high temperatures, which makes it more difficult to dissipate the heat.

In applications where it is desirable to have reference voltages larger than 1.24 V, a modified bandgap reference as shown in Fig. 7.11 can be used. It is not difficult to show that the output voltage is now given by

$$V_{ref\text{-}0} = \left(1 + \frac{R_4}{R_5}\right)\left[V_{G0} + (m-1)\frac{kT_0}{q}\right] \cong \left(1 + \frac{R_4}{R_5}\right)1.24 \text{ V} \tag{7.38}$$

Resistor R_3 has been added to cancel the effects of the finite base currents going through R_4 and should be chosen according to the formula in the figure. The interested reader is referred to [Brokaw, 1974] for additional details concerning this realization.

CMOS Bandgap References

The most popular method for realizing CMOS voltage references also makes use of a bandgap voltage reference despite the fact that *independent* bipolar transistors are not available. These CMOS circuits rely on using what are commonly called *well transistors.* These devices are vertical bipolar transistors that use wells as their bases and the substrate as their collectors. In an n-well process (a common modern process), these vertical bipolar transistors are pnp types with their collectors connected to ground, as shown in Fig. 7.12(*a*). In a p-well process, they would be npn transistors with their collectors connected to the positive power supply, as

> ***Key Point:*** *The implementation of bandgap references in a CMOS process makes use of bipolar transistors formed using doped well regions in the silicon substrate.*

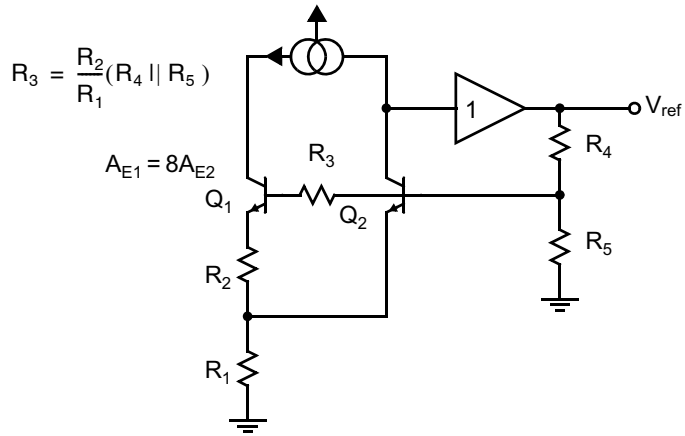


Fig. 7.11 A bipolar bandgap with output voltages greater than 1.24 V.

shown in Fig. 7.12(*b*). These transistors have reasonable current gains, but their main limitation is the series base resistance, which can be high due to the large lateral dimensions between the base contact and the effective emitter region. To minimize errors due to this base resistance, the maximum collector currents through the transistors are usually constrained to be less than 0.1 mA. It is possible to use these transistors to implement bandgap voltage references using configurations similar to those shown in Fig. 7.13(*a*) for n-well processes [Kujik, 1973] or Fig. 7.13(*b*) for p-well processes [Ye, 1982].

With respect to the n-well implementation of Fig. 7.13(*a*), we have

$$V_{ref} = V_{EB1} + V_{R1} \tag{7.39}$$

Also, assuming the opamp has large gain and that its input terminals are at the same voltage, then

$$V_{R2} = V_{EB1} - V_{EB2} = \Delta V_{EB} \tag{7.40}$$

Now, since the current through R_3 is the same as the current through R_2, we have

$$V_{R3} = \frac{R_3}{R_2} V_{R2} = \frac{R_3}{R_2} \Delta V_{EB} \tag{7.41}$$

using (7.40). The opamp feedback also makes the voltage across R_1 equal to the voltage across R_3. Using this fact and substituting (7.41) into (7.39) results in

$$V_{ref} = V_{EB1} + \frac{R_3}{R_2} \Delta V_{EB} \tag{7.42}$$

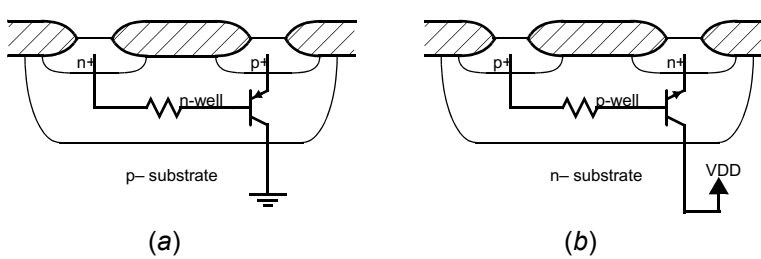


Fig. 7.12 Vertical CMOS well transistors realized in (*a*) an n-well process and (*b*) a p-well process.

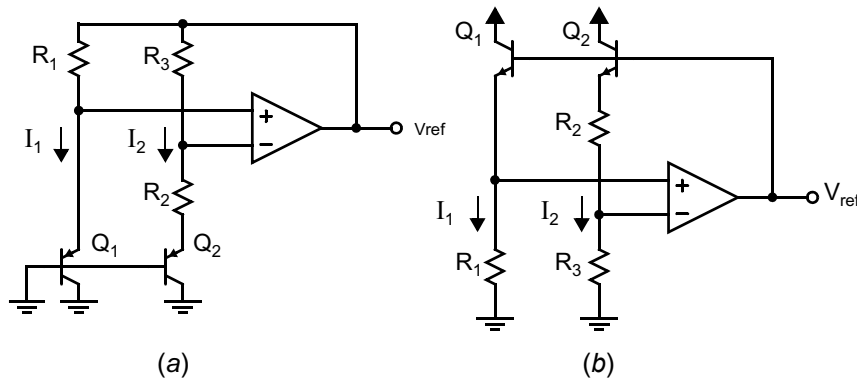


Fig. 7.13 Bandgap voltage references implemented with well transistors in (*a*) an n-well CMOS process, and (*b*) a p-well CMOS process.

which is in the required form to realize a bandgap reference. In integrated realizations of this reference, the bipolar transistors are often taken the same size, and the different current-densities are realized by taking R_3 greater than R_1, which causes I_1 to be larger than I_2. In this case, we would have

$$\frac{J_1}{J_2} = \frac{R_3}{R_1} \tag{7.43}$$

since R_1 and R_3 have the same voltage across them. Also, recalling from (7.16) that

$$\Delta V_{EB} = V_{EB1} - V_{EB2} = \frac{kT}{q} \ln\left(\frac{J_1}{J_2}\right) \tag{7.44}$$

and using (7.43) in (7.42) gives

$$V_{ref} = V_{EB1} + \frac{R_3}{R_2}\frac{kT}{q} \ln\left(\frac{R_3}{R_1}\right) \tag{7.45}$$

It is immediately recognizable that

$$K = \frac{R_3}{R_2} \tag{7.46}$$

EXAMPLE 7.5

Find the resistances of a bandgap voltage reference based on Fig. 7.13(*a*), where $I_1 = 80\ \mu A$, $I_2 = 8\ \mu A$, and $V_{EB1\text{-}0} = 0.65$ V at $T = 300$ K.

Solution

Recalling from (7.24) that

$$V_{ref\text{-}0} = 1.24 \text{ V} \tag{7.47}$$

therefore, from (7.39), we require

$$V_{R1} = V_{ref\text{-}0} - V_{EB1\text{-}0} = 0.59 \text{ V} \tag{7.48}$$

Also, since $V_{R3} = V_{R1}$, we have

$$R_3 = \frac{V_{R3}}{I_2} = \frac{0.59\ \text{V}}{8\ \mu\text{A}} = 73.8\ \text{k}\Omega \tag{7.49}$$

and

$$R_1 = R_3\frac{I_2}{I_1} = 7.38\ \text{k}\Omega \tag{7.50}$$

Now, recalling from (7.26) that

$$K = \frac{1.24 - 0.65\ \text{V}}{0.0258 \times \ln(10)} = 9.93 \tag{7.51}$$

therefore

$$R_2 = \frac{R_3}{K} = 7.43\ \text{k}\Omega \tag{7.52}$$

It is of interest to note here that using (7.44) and noting that $J_1/J_2 = I_1/I_2$ (since the sizes of Q_1 and Q_2 are assumed to be the same), we find that

$$\Delta V_{EB} = \frac{kT_0}{q}\ln\left(\frac{I_1}{I_2}\right) = 59\ \text{mV} \tag{7.53}$$

gives the temperature dependence value of $0.198\ \text{mV/K}$, as found in Example 7.3, which requires approximately a gain of 10 to cancel temperature dependency (actually 9.93 in this case).

The design equations for a voltage reference that is suitable for p-well processes and shown in Fig. 7.13(*b*) are essentially identical to those just given for the n-well reference.

In CMOS realizations of the references just described, the large value resistors are often realized by well resistors. Unfortunately, these resistors have a temperature dependence approximately given by [Michejda, 1984]

$$R(T) = R_0\frac{T^\eta}{T_0^\eta} \tag{7.54}$$

where $\eta = 2.2$. The errors caused by this temperature dependence can be eliminated by offsetting $V_{\text{ref-0}}$ slightly positive from the value given by (7.23) to

$$V_{\text{ref-0}} = V_{G0} + (m + \eta - 1)\frac{kT_0}{q} \tag{7.55}$$

Assuming the effects of the temperature coefficient of the resistors have been minimized, the next major source of error is often due to the input-offset voltage of the opamp [Michejda, 1984]. This results in an error term in the equation for ΔV_{BE} that is roughly equal to K times the input-offset voltage of the opamp. For example, a 1mV offset error that is temperature independent causes a temperature coefficient (TC), error approximately given by [Song, 1983]

$$TC_{\text{error}} \cong 26\ \text{ppm/°C} \tag{7.56}$$

One means of eliminating this source of error is to use switched-capacitor (SC) amplifiers that have input-offset compensation circuits [Song, 1983]. One possible SC-based voltage reference is shown in Fig. 7.14, where it makes use of an amplifier described in [Martin, 1987]. The amplifier shown results in a circuit having its output

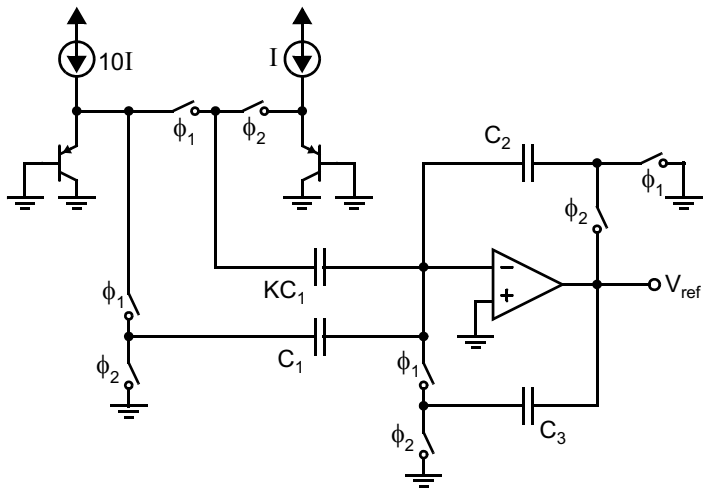


Fig. 7.14 An SC-based voltage reference that is insensitive to opamp input-offset voltages.

valid at all times and is less sensitive to finite opamp gain. A detailed explanation of how the amplifier shown in Fig. 7.14 operates is deferred until Chapter 14, where switched-capacitor circuits are discussed.

Assuming errors due to the input-offset voltage of the opamp have been minimized, there is still an inherent temperature dependence of the bandgap voltage reference, as we saw in (7.27). In addition, a further temperature dependence occurs because V_{G0} varies slightly with temperature (which has been ignored in the above analysis). Together, these two error sources limit the best achievable temperature coefficient to about 25 ppm/°K. Minimizing these second-order effects is beyond the scope of this book, but the interested reader is referred to [Palmer, 1981; Meijer, 1982; Song, 1983] to see examples of how errors due to second-order effects have been minimized.

An alternative realization of a CMOS bandgap reference was reported in [Degrauwe, 1985], where lateral npn well transistors were used. Also, in [Tzanateas, 1979] a voltage reference was reported that was based on a realized PTAT voltage from the difference of the source-gate voltages of two MOS transistors biased in weak inversion. The interested reader can consult the references for details on these circuits.

7.3.3 Low-Voltage Bandgap Reference

A problem with the bandgap circuits described so far is that they are not compatible with supply voltages of 1.5 V or less. The fundamental problem is that all these circuits produce, in series, the inverse-PTAT voltage across a pn-junction and a PTAT voltage scaled to provide temperature-independence. Together, these voltages exceed 1 V making it impossible to use these circuits with a 1-V supply. The solution is to instead sum *currents*: one inverse-PTAT and one PTAT, passing them through a resistor to form a temperature-insensitive reference voltage below 1 V.

Key Point: *Connecting PTAT and IPTAT voltages in series to produce a precision reference results in a voltage larger than the bandgap of silicon. To produce a bandgap reference under a lower supply voltage, it is necessary to first convert the PTAT and IPTAT quantities into currents, and then sum them by shunting.*

A low-voltage bandgap circuit based on this idea is shown in Fig. 7.15 [Banba, 1999]. The feedback loop formed by the amplifier and current mirror Q_1/Q_2 ensures that the voltages at the drains of Q_1 and Q_2 are very well matched and equal to the forward voltage drop across diode D_1, $V_{BE,1}$, which is inversely proportional to absolute temperature.

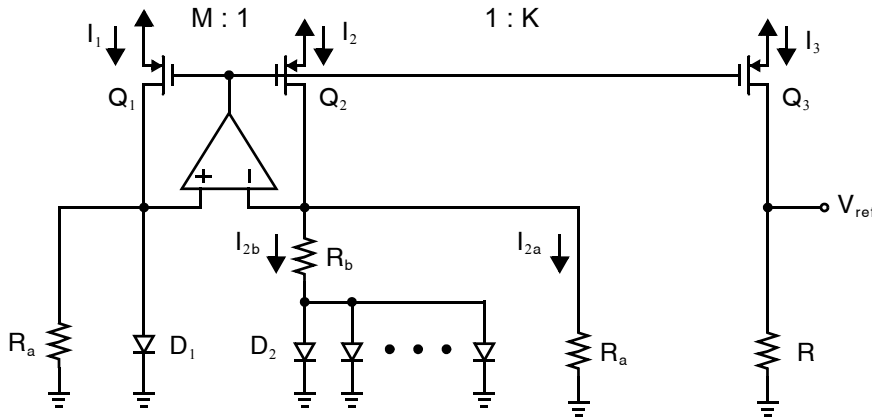


Fig. 7.15 A low-voltage bandgap circuit [Banba, 1999].

The voltage across R_b is equal to the difference between the forward bias voltages of D_1 and D_2, ΔV_{BE}, and is therefore PTAT. The current mirror and matched resistors R_a ensure that the current through the diodes differs by the factor M, and that the diode current densities differ by a factor MN.

$$\Delta V_{BE} = \frac{kT}{q}\ln(MN) \tag{7.57}$$

The inverse PTAT $V_{BE,1}$ and PTAT ΔV_{BE} are converted into the currents I_{2a} and I_{2b} by the resistors R_a and R_b, respectively.

$$I_{2a} = \frac{V_{BE,1}}{R_a} \tag{7.58}$$

$$I_{2b} = \frac{\Delta V_{BE}}{R_b} \tag{7.59}$$

These two currents are summed at the drain of Q_2 and mirrored to device Q_3. The resulting current then passes back through resistor R to yield the reference output.

$$V_{ref} = KI_2R = V_{BE,1}\left(\frac{KR}{R_a}\right) + \frac{kT}{q}\ln(MN)\left(\frac{KR}{R_b}\right) \tag{7.60}$$

The values M, N, R_a, and R_b are chosen to ensure that the inverse temperature coefficient of the first term is precisely cancelled by the positive temperature coefficient of the second term. Notice that nowhere are there DC voltages exceeding V_{BE}, so this circuit has been used with supply voltages as low as 0.8 V. The values of K and R can be chosen to scale the actual value of the reference potential at V_{ref}, while the values of N, R_a and R_b are chosen to cancel the negative and positive temperature coefficients of the first and second terms in (7.60). Opamp offset and mismatch in the current mirror devices, diodes, and resistors are all sources of error.

7.3.4 Current Reference

Once a bandgap voltage has been established, it can be used in combination with a resistor to provide a reference current. The basic approach is illustrated in Fig. 7.16. Feedback is used to generate the PMOS gate voltage that results in a voltage drop of precisely V_{ref} across a resistor R. This gate voltage can then be used to generate a scaled copy of the resulting current, $I_{ref} = MV_{ref}/R$. If R is an on-chip resistor, the reference current will vary

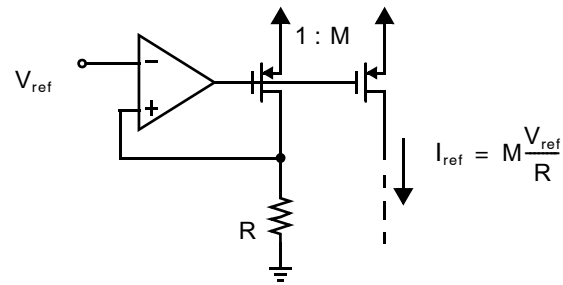


Fig. 7.16 Obtaining a reference current from a voltage reference.

with the value of on-chip resistors. This may be desirable if the current is passing through another on-chip resistor matched to R, in which case the resulting voltage drop will be stable. If, however, an absolute current reference is required, a precision resistor may be realized either off-chip or by trimming R and somehow compensating for its temperature variations.

7.4 VOLTAGE REGULATION

A voltage regulator produces a low-noise dc voltage from which some current can be drawn. Its primary use is to provide a quiet supply voltage to an analog circuit, particularly in environments when a noisy supply voltage will otherwise limit performance.

A basic voltage regulator is shown in Fig. 7.17. It accepts as input a reference voltage, V_{ref}, and produces as output V_{reg}. It is essentially a unity-gain feedback buffer. The amplifier provides gain in the loop which ensures $V_{reg} \cong V_{ref}$ while Q_1, called the pass transistor, sources the load current. The reference input may be derived from a bandgap circuit, if it is to remain fixed over process and temperature variations. Variable-output regulators have a fixed reference voltage and incorporate a resistive voltage divider in the feedback loop which is adjusted to effect change in the output voltage.

Key Point: *A voltage regulator comprises a feedback loop that keeps an output voltage equal to a reference voltage by controlling the flow of current to the output through a pass-transistor.*

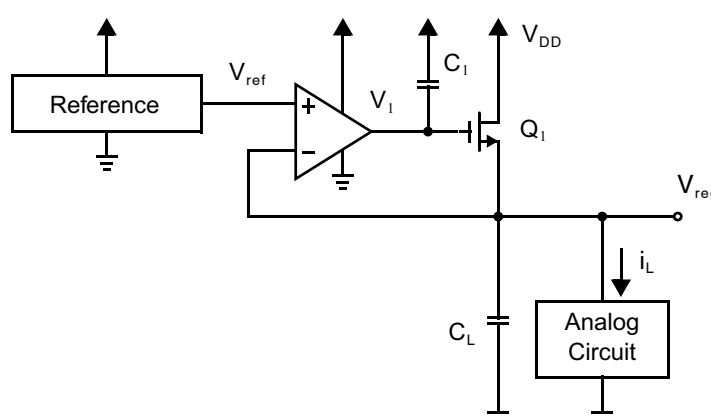


Fig. 7.17 A basic voltage regulator.

7.4.1 Regulator Specifications

Power Supply Rejection

A regulator's ability to maintain a quiet voltage at V_{reg} in the presence of noise on V_{DD} is measured by superimposing a small signal onto V_{DD} and measuring the resulting variations in V_{reg}. The ratio of the two small signals v_{dd} and v_{reg} is its power supply rejection ratio which is usually expressed in decibels and is frequency-dependent.

$$PSRR(\omega) = 20\log_{10}|v_{dd}/v_{reg}| \text{ dB} \tag{7.61}$$

Evaluated at dc, the inverse of supply rejection may be referred to as the *line regulation*.

Clearly the supply rejection of the reference generator and the opamp play important roles in the supply rejection of a regulator. For example, since V_{reg} will track V_{ref} within the feedback loop's bandwidth, low frequency supply rejection is often determined by the supply rejection of the reference generator.

Output Impedance

Variations in the current drawn by the regulator's load, i_L, will generally cause variations in V_{reg}. If the changes are small, they are related by the regulator's small-signal output impedance, Z_{out}. At high frequencies, this will be determined by the impedance of the load capacitance, $1/j\omega C_L$. At low frequencies, it is determined by the impedance of the pass transistor $1/g_{m,1}$ divided by the opamp gain.

Dropout Voltage

An important specification of analog integrated circuit regulators is the maximum regulated voltage that can be maintained under a given global supply voltage, or equivalently the minimum voltage drop between the global supply and the regulated voltage, $V_{DD} - V_{reg}$, called the *dropout voltage*, V_{DO}. Notice that the dc power dissipation in the transistor Q_1 is

$$P_1 = (V_{DD} - V_{reg})i_L \tag{7.62}$$

This power is not being delivered to the load, so it is a major inefficiency in the system. A lower limit on this power is imposed by the dropout voltage,

$$P_{1,min} = V_{DO}i_L \tag{7.63}$$

motivating one to operate with $(V_{DD} - V_{reg})$ near V_{DO} and to minimize V_{DO}.

Looking at Fig. 7.17, clearly a voltage of $V_{eff,1}$ must be maintained between the drain and source of Q_1, otherwise the transconductance of Q_1 will drop, causing a decrease in loop gain and in the accuracy with which V_{reg} tracks V_{ref}. A more important consideration in determining the dropout voltage, however, is that V_{reg} must be below V_1 by at least $V_{GS,1}$. If the opamp is operated under the same supply voltage as Q_1, and assuming the opamp output must be at least V_{eff} below V_{DD} to prevent some device there from entering triode, the dropout voltage is at least $2V_{eff} + V_{tn,1}$. This quantity is not too large if native NMOS devices (having $V_{tn} \approx 0$) are available. Alternatively, in some cases the opamp is operated from a higher supply voltage.

7.4.2 Feedback Analysis

The open loop analysis of the voltage regulator with NMOS pass device in Fig. 7.17 is very similar to that of a basic source follower. Assuming a single-stage opamp with transconductance G_{ma} and output resistance R_{oa}, the loop is broken at the opamp's input and a test signal is injected, v_t. The resulting small-signal equivalent circuit is shown in Fig. 7.18. The regulator load is modeled by the small-signal resistance R_L and combined with r_{ds1} and $1/g_{s1}$ to form R_L'. Applying the results from Chapter 4, the open-loop response is

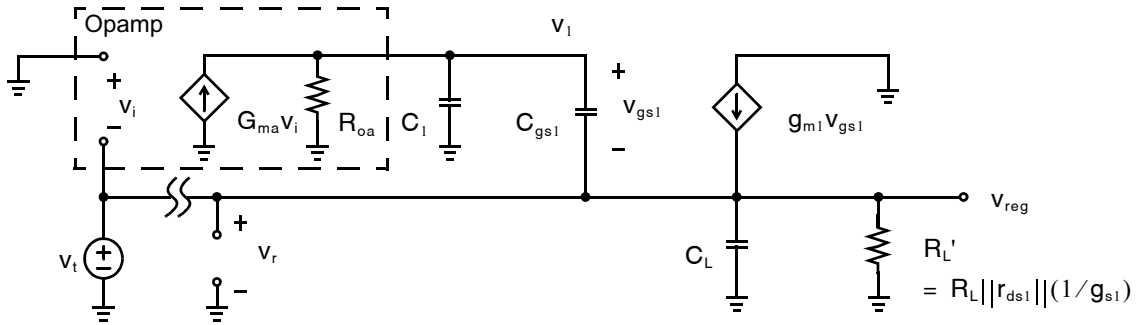


Fig. 7.18 The open-loop small-signal schematic of a voltage regulator with NMOS pass device.

$$L(s) = \left(\frac{G_{ma}R_{oa}R_L'}{R_L' + 1/g_{m1}}\right)\frac{\left(1 + \dfrac{sC_{gs1}}{g_{m1}}\right)}{\left(1 + \dfrac{s}{\omega_0 Q} + \dfrac{s^2}{\omega_0^2}\right)} \tag{7.64}$$

where

$$\omega_0 = \sqrt{\frac{(1/R_{oa})(g_{m1} + 1/R_L')}{C_{gs1}C_L + C_1(C_{gs1} + C_L)}} \tag{7.65}$$

$$Q = \frac{\sqrt{(1/R_{oa})(g_{m1} + 1/R_L')[C_{gs1}C_L + C_1(C_{gs1} + C_L)]}}{C_L/R_{oa} + C_1(g_{m1} + 1/R_{oa}) + C_{gs1}/R_L'} \tag{7.66}$$

Assuming the loop is dominant-pole compensated, then $Q \ll 1$ and the denominator (7.64) may be factored into two first-order terms.

$$L(s) = \left(\frac{G_{ma}R_{oa}R_L'}{R_L' + 1/g_{m1}}\right)\frac{\left(1 + \dfrac{s}{\omega_z}\right)}{\left(1 + \dfrac{s}{\omega_{p1}}\right)\left(1 + \dfrac{s}{\omega_{p2}}\right)} \tag{7.67}$$

Hence, there are poles at

$$-\omega_{p1} = -\omega_0 Q = -\frac{(1/R_{oa})(g_{m1} + 1/R_L')}{C_L/R_{oa} + C_1(g_{m1} + 1/R_{oa}) + C_{gs1}/R_L'} \tag{7.68}$$

$$-\omega_{p2} = -\frac{\omega_0}{Q} = -\frac{C_L/R_{oa} + C_1(g_{m1} + 1/R_{oa}) + C_{gs1}/R_L'}{C_{gs1}C_L + C_1(C_{gs1} + C_L)} \tag{7.69}$$

and there is a zero due to the feedforward signal path through C_{gs1}

$$-\omega_z = -g_{m1}/C_{gs1} \tag{7.70}$$

There are two different ways to compensate the loop. If C_1 is made large by including an additional compensation capacitor between V_1 and ground, $C_1 g_{m1}$ becomes the dominant term in the denominator of (7.68) and in the numerator of (7.69). As a result, the dominant pole $\omega_{p1} \approx 1/R_{oa}C_1$ is associated with the time constant at

v_1 and $\omega_{p2} \approx g_{m1}/C_L$ is associated with the time constant at v_{reg}. In this case, a capacitor is still required at the output C_L to filter high frequency noise on v_{reg}. Alternatively, C_L can be made very large and R_{oa} decreased so that C_L/R_{oa} is a dominant term in both (7.68) and (7.69). In this case, the two poles change roles so that ω_{p1} is the output pole and ω_{p2} is the pole at v_1. This generally requires a larger compensation capacitor and/or higher power consumption in the opamp, but it provides a large capacitance at the output where it also improves supply rejection and output impedance at medium-to-high frequencies. Generally, ω_z will be at a much higher frequency than the poles and have little effect on stability.

7.4.3 Low Dropout Regulators

When the regulated output must be at a voltage only 200–400 mV below V_{DD}, and especially when native (near-zero-V_t) NMOS devices are unavailable, it is necessary to utilize a PMOS device for Q_1 as shown in Fig. 7.19. In this case, the gate voltage V_1 is well below V_{DD} so the dropout voltage is only limited by $V_{eff,1}$. This is commonly referred to as a *low dropout (LDO) voltage regulator* and is popular when power efficiency is critical. Notice that the polarity of the opamp's input terminals has been reversed because there is now negative small-signal gain from V_1 to V_{reg}. Although the resistance seen looking into Q_1 is increased compared with Fig. 7.17 from $1/g_{m,1}$ to $r_{ds,1}$, the loop gain is also increased by roughly a factor $g_{m,1}r_{ds,1}$ so the regulator's closed-loop output resistance is roughly the same.

Key Point: *Whereas an NMOS pass-transistor provides superior supply rejection, unless native NMOS devices are available, a PMOS pass-transistor permits a lower dropout voltage and, hence, better efficiency.*

A significant drawback of PMOS devices is their reduced power supply rejection. With respect to small signals at V_{DD}, Q_1 behaves like a common-gate amplifier having considerable gain. At low frequencies, the feedback loop tracks and cancels the resulting variations in V_{reg}, and at high frequencies the variations are filtered by C_L, but there is generally a mid-band frequency range where it is difficult to provide good supply rejection using a LDO.

The feedback analysis of a LDO is similar to that of a two-stage opamp, where the opamp in Fig. 7.19 is presumed to have only a single stage and the common-source pass transistor Q_1 provides the second gain stage. Neglecting C_{gd1}, there are clearly two poles in the open loop response corresponding to the two nodes in the system.

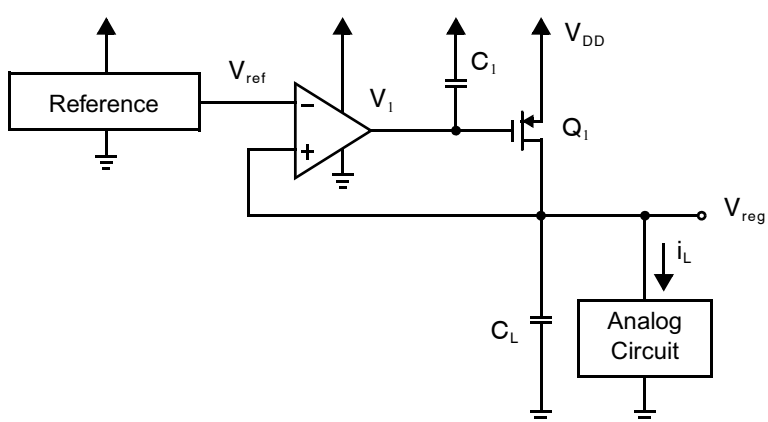


Fig. 7.19 A low dropout (LDO) voltage regulator.

$$L(s) = \frac{G_{ma}R_{oa}g_{m1}R_L'}{\left(1+\frac{s}{\omega_{pa}}\right)\left(1+\frac{s}{\omega_{pL}}\right)} \tag{7.71}$$

The pole at the amplifier output (gate of Q_1) is

$$\omega_{pa} = \frac{1}{R_{oa}C_1'} \tag{7.72}$$

where $C_1' = C_1 + C_{gs1}$, and the output pole is

$$\omega_{pL} = \frac{1}{R_L'C_L} \tag{7.73}$$

Again, either pole can be made dominant to compensate the loop. A dominant output pole, $\omega_{pL} \ll \omega_{pa}$, requires high power consumption in the opamp to keep R_{oa} low, but has the advantage of providing a low output impedance at medium frequencies which improves the supply rejection. This is particularly important in LDOs since the PMOS pass transistor tends to amplify supply noise as described above.

Key Point: *Compensation of a voltage regulator presents a difficult decision: either make the output pole dominant, resulting in high power consumption in the feedback amplifier, or make the pass-transistor gate node dominant, in which case supply rejection is worse.*

EXAMPLE 7.6

What is the supply rejection of the basic voltage regulator shown in Fig. 7.20?

Solution

A small-signal equivalent circuit is shown in Fig. 7.21 with a small-signal source shown at V_{DD}. With the loop open, the response from v_{DD} to v_{reg} is simply that of a common-gate amplifier loaded by $R_L||(1/sC_L)$.

$$H_{CG}(s) = \left(\frac{R_L}{r_{ds1}+R_L}\right)\frac{1}{1+sR_LC_L} \tag{7.74}$$

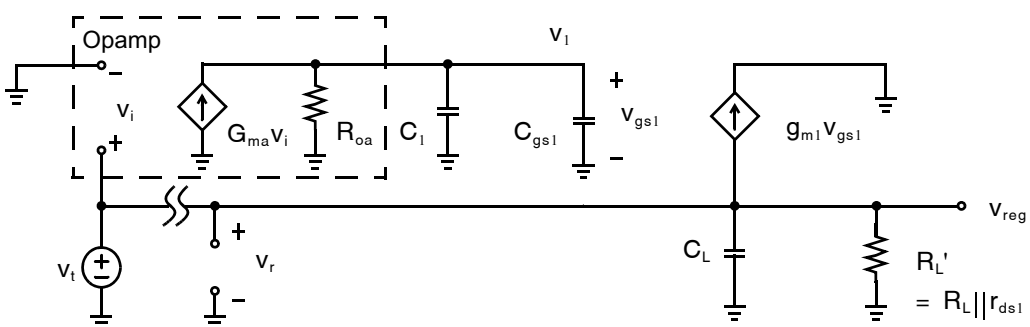


Fig. 7.20 The open-loop small-signal schematic of a LDO voltage regulator with a PMOS pass device.

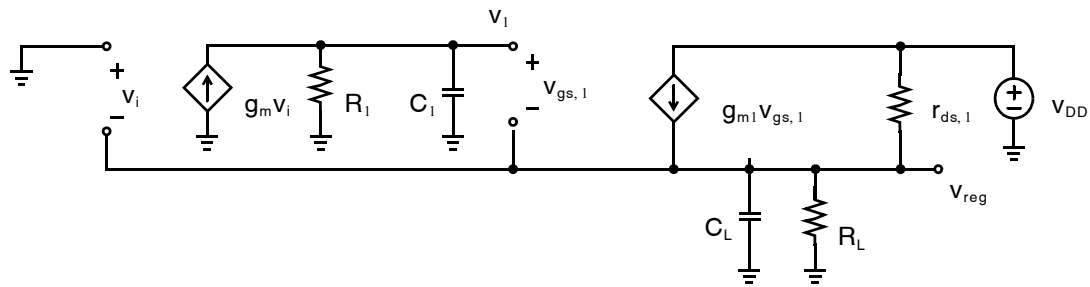


Fig. 7.21 The small-signal schematic of a basic voltage regulator.

With the loop closed, this first-order lowpass response is shaped by $1/(1 + L(s))$. The resulting frequency response from small signals on the supply voltage to small signals at the regulator output is the inverse of the power supply rejection

$$\frac{v_{reg}}{v_{dd}}(s) = PSRR^{-1}(s) = \frac{H_{CG}(s)}{1 + L(s)} \tag{7.75}$$

The responses are plotted in Fig. 7.22 for two cases: where the load pole ω_{pL} is dominant, Fig. 7.22(*c*), and where the opamp output pole ω_{pa} is dominant, Fig. 7.22(*d*). In both cases, the dc rejection (line regulation) is given by

$$PSRR^{-1}(s) = \left(\frac{R_L}{r_{ds1} + R_L}\right)\frac{1}{G_{ma}R_{oa}g_{m1}R_L'} \tag{7.76}$$

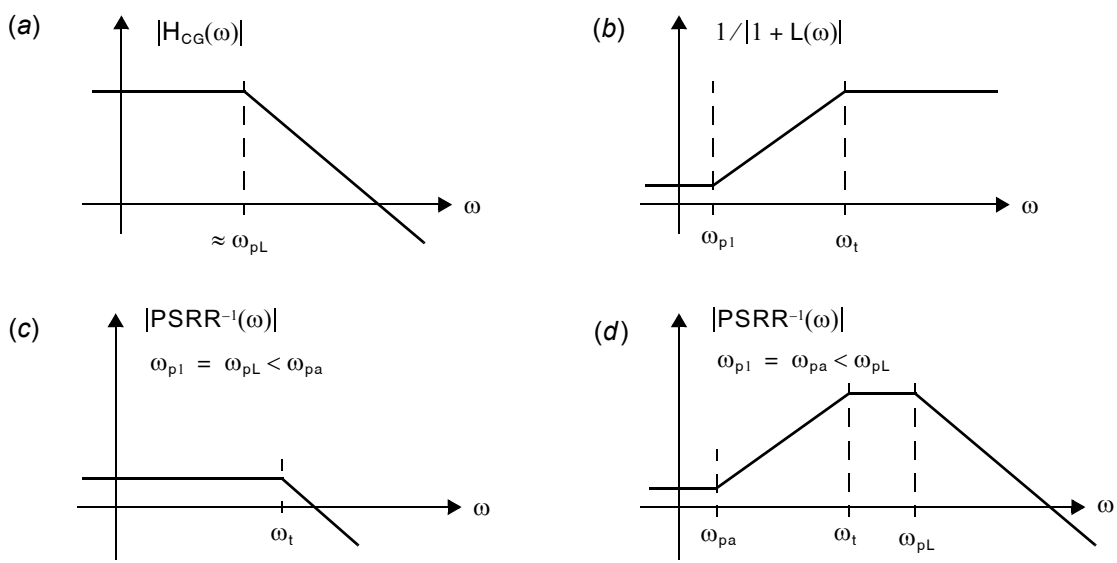


Fig. 7.22 The power supply rejection of a LDO regulator: (*a*) open loop supply rejection; (*b*) closed-loop shaping frequency response; (*c*) $|PSRR(\omega)| = |H_{CG}(\omega)|/|1 + L(\omega)|$ with $\omega_{pL} < \omega_{pa}$; (*d*) $|PSRR(\omega)| = |H_{CG}(\omega)|/|1 + L(\omega)|$ with $\omega_{pa} < \omega_{pL}$.

As the loop gain increases beyond its dominant pole frequency ω_{p1}, the supply rejection worsens. In the first case, this occurs at the same frequency as the output pole begins to filter out the supply noise, so the supply rejection remains high. In the second case, the opamp pole ω_{pa} occurs first at frequencies where the load capacitor is still unable to provide filtering of supply-induced noise, so the supply rejection worsens until the load pole kicks in. Unfortunately, for stability the poles ω_{pa} and ω_{pL} must be widely spaced, so the deterioration is significant. This usually occurs at frequencies in the range of 100's of MHz where supply noise due to digital circuitry can be very high, and is therefore a major challenge.

7.5 SUMMARY OF KEY POINTS

- A bias circuit generates the dc voltages required to keep transistors near some desired operating point; as transistor parameters change, either from chip to chip or with changes in temperature, so must the bias voltages. [p. 303]
- A reference circuit generates a voltage and/or current of a known fixed absolute value. [p. 305]
- A regulator circuit improves the quality of a dc voltage or current, usually decreasing the noise. [p. 306]
- A constant-transconductance bias circuit produces a small-signal transconductance that is a fixed fraction of a resistor value, independent of process, temperature, or supply variations. [p. 308]
- The constant-transconductance bias circuit ensures a constant transconductance in all transistors of the same type as Q_{13}. However, the transconductance of transistors of the opposite type will vary. [p. 308]
- Many bias circuits require auxiliary start-up circuits to prevent them from resting in a zero-current state indefinitely. [p. 309]
- An integrated voltage reference is made by adding the forward bias voltage of a pn junction to the difference in the forward voltages of two pn junctions biased at different current densities. Before adding them, these voltages are scaled so that their respective positive and negative temperature coefficients cancel precisely. [p. 311]
- The implementation of bandgap references in a CMOS process makes use of bipolar transistors formed using doped well regions in the silicon substrate. [p. 315]
- Connecting PTAT and IPTAT voltages in series to produce a precision reference results in a voltage larger than the bandgap of silicon. To produce a bandgap reference under a lower supply voltage, it is necessary to first convert the PTAT and IPTAT quantities into currents, and then sum them by shunting. [p. 319]
- A voltage regulator comprises a feedback loop that keeps an output voltage equal to a reference voltage by controlling the flow of current to the output through a pass-transistor. [p. 321]
- Whereas an NMOS pass-transistor provides superior supply rejection, unless native NMOS devices are available, a PMOS pass-transistor permits a lower dropout voltage and, hence, better efficiency. [p. 324]
- Compensation of a voltage regulator presents a difficult decision: either make the output pole dominant, resulting in high power consumption in the feedback amplifier, or make the pass-transistor gate node dominant, in which case supply rejection is worse. [p. 325]

7.6 REFERENCES

H. Banba, H. Shiga, A. Umezawa, T. Miyaba, T. Tanzawa, S. Atsumi, and K. Sakui. "A CMOS Bandgap Reference Circuit with Sub-1-V Operation," *IEEE Journal of Solid-State Circuits*, Vol. 34, no. 5, pp. 670–674, May 1999.

P. Brokaw. "A Simple Three-Terminal IC Bandgap Reference," *IEEE Journal of Solid-State Circuits,* Vol. SC-9, pp. 388–393, December 1974.

J. Brugler. "Silicon Transistor Biasing for Linear Collector Current Temperature Dependence," *IEEE Journal of Solid-State Circuits,* Vol. SC-2, pp. 57–58, June 1967.

M. Degrauwe, O. Leuthold, E. Vittoz, H. Oguey, and A. Descombes, "CMOS Voltage References Using Lateral Bipolar Transistors," *IEEE Journal of Solid-State Circuits,* Vol. 20, pp. 1151–1157, December 1985.

K. Kujik. "A Precision Reference Voltage Source," *IEEE Journal of Solid-State Circuits*, Vol. SC-8, pp. 222–226, June 1973.

K. Martin, L. Ozcolak, Y. S. Lee, and G. C. Temes. "A Differential Switched-Capacitor Amplifier," *IEEE Journal of Solid-State Circuits,* Vol. SC-22, no. 1, pp. 104–106, February 1987.

A. McLaren and K. Martin. "Generation of Accurate On-Chip Time Constants and Stable Transconductances," *IEEE Journal of Solid-State Circuits*, Vol. 36, no. 4, pp. 691–695, April 2001.

G. Meijer, P. Schmale, and K. van Zalinge, "A New Curvature-Corrected Bandgap Reference," *IEEE Journal of Solid-State Circuits,* Vol. SC-17, pp. 1139–1143, December 1982.

J. Michejda and S. Kim, "A Precision CMOS Bandgap Reference," *IEEE J. of Solid-State Circuits,* Vol. SC-19, no. 6, pp. 1014–1021, December 1984.

S. Nicolson and K. Phang, "Improvements in Biasing and Compensation of CMOS Opamps," *IEEE International Symposium on Circuits and Systems*, Vol. 1, pp. 665–668, May 2004.

C. Palmer and R. Dobkin, "A Curvature-Corrected Micropower Voltage Reference," *IEEE International Solid-State Circuits Conf.*, pp. 58–59, February 1981.

B. Song and P. Gray, "A Precision Curvature-Compensated CMOS Bandgap Reference," *IEEE Journal of Solid-State Circuits,* Vol. SC-18, no. 6, pp. 634–643, December 1983.

J. M. Steininger, "Understanding Wide-Band MOS Transistors," *IEEE Circuits and Devices*, Vol. 6, No. 3, pp. 26–31, May 1990.

Y. Tsividis, "Accurate Analysis of Temperature Effects in I_C-V_{BE} Characteristics with Application to Bandgap Reference Sources," *IEEE Journal of Solid-State Circuits,* Vol. 15, no. 6, pp. 1076–1084, December 1980.

G. Tzanateas, C. Salama, and Y. Tsividis, "A CMOS Bandgap Voltage Reference," *IEEE Journal of Solid-State Circuits,* Vol. SC-14, pp. 655–657, June 1979.

R. Ye and Y. Tsividis, "Bandgap Voltage Reference Sources in CMOS Technology," *Electron. Lett.,* Vol. 18, no. 1, January 1982.

7.7 PROBLEMS

7.7.1 Section 7.1: Analog Integrated Circuit Biasing

SPICE! *Most of these problems can be performed with the aid of Spice simulations using the provided device models.*

7.1 A NMOS triode transistor with $(W/L) = 20$ in the 0.18-μm technology of Table 1.5 has $V_S = V_D = V_B = 0.3$ V.

a. Find V_G so that $r_{ds} = 1$ kΩ.

b. What should the new value of V_G be if μ_n increases by 30%, but r_{ds} remains 1 kΩ?

7.2 The circuit of Fig. P7.2 is to be realized in the 0.35-μm technology of Table 1.5 with $L_{1,2} = 0.4$ μm.

a. Find R, W_1, and W_2 so that $V_{eff1,2} = 300$ mV and $I_D = 50$ μA.

b. How much will I_D change if both V_{tn} and $|V_{tp}|$ increase by 25%?

c. How much will the transconductance change if R decreases by 15%?

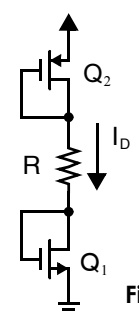


Fig. P7.2

7.3 Assume a constant V_b in Example 7.1. How much does the gain vary under a 10% decrease in $\mu_n C_{ox}$, and a 10% increase in both V_{tn} and R?

7.7.2 Section 7.2: Establishing Constant Transconductance

7.4 Using the 0.35-μm device parameters in Table 1.5, design the bias circuit of Fig. 7.6 to have $V_{eff} = 0.25$ V at 25°C for all transistors except Q_{15}. For $Q_{10} - Q_{14}$ having sizes of 4 μm/0.3 μm and Q_{15} having a size of 16 μm/0.3 μm, what is the required value of R_B? What would the effective gate-source voltage be at 70°C assuming electron mobility varies proportional to $T^{-3/2}$?

7.5 Assuming that μ_n is proportional to $T^{-3/2}$, what would V_{eff13} be in Fig. 7.6 at 100 °C if it was 0.2 V at 20°C?

7.6 Assume that R_B from Fig. 7.6 has a temperature dependence of +0.3%/°C. What would V_{eff13} be at 100 °C if it was 0.2 V at 20 °C? Assume μ_n varies as in Problem 7.5.

7.7 Using the 0.18-μm devices in Table 1.5, what is the value required for R_B in Fig. 7.7 to give $V_{eff14} = 0.2$ V? Assume $(W/L)_{12,13} = (2\ \mu m/0.2\ \mu m)$ and $(W/L)_{12,13} = (8\ \mu m/0.2\ \mu m)$.

7.8 Size the devices in Fig. 7.7 so that resistor $R_B = 1\ k\Omega$ has a current of 120 μA flowing through it. All gate lengths are $L = 0.25\ \mu m$ in the 0.18-μm technology of Table 1.5.

7.9 Repeat Problem 7.8 for using the 0.35-μm device parameters in Table 1.5.

7.10 What is the effect of an dc input offset of V_{OS} for the opamp in Fig. 7.7? Does the circuit still provide a constant transconductance?

7.11 A PMOS differential pair is to be biased to have a constant transconductance of $g_m = 1$ mA/V. Design the bias circuit based of Fig. 7.7 to achieve this in the 0.18-μm technology of Table 1.5. The power consumption of the bias circuit (neglecting the opamp) should be 100 μW.

7.7.3 Section 7.3: Establishing Constant Voltages and Currents

7.12 Assume two bipolar transistors in a PTAT are biased at a current-density ratio of 8 : 1 at $T = 320°K$. What is the difference in their base-emitter voltages and what is its temperature dependence?

7.13 Assuming $V_{BE0\text{-}2} = 0.65$ V and using (7.25), what is the value required for K in order to get zero temperature dependence at $T = 320$ °K for a current-density ratio of 8:1 in the two transistors?

7.14 Prove equations (7.26) and (7.27).

7.15 Find values for the bipolar reference of Fig. 7.11, where $V_{ref} = 2.5$ V.

7.16 Prove that equation (7.45) holds for the voltage reference of Fig. 7.13(*b*).

7.17 Consider the circuit in Fig. P7.17.

a. A start-up circuit is needed for the bias circuit. What are the startup node voltages that will cause the circuit not to startup properly?

b. Find an expression for V_{out} in terms of the circuit parameters. Under what conditions is a bandgap reference? You may assume infinite r_{ds} for the MOSFETs in active mode and no mismatch.

c. Select values for n, R_1, and R_2 to make this circuit a bandgap reference, assuming V_{EB} of the large BJTs has a temperature coefficient of -2 mV/K.

d. If the threshold voltages of the two NMOS devices differ by ΔV_{tn}, give an expression for V_{out} in terms of ΔV_{tn}.

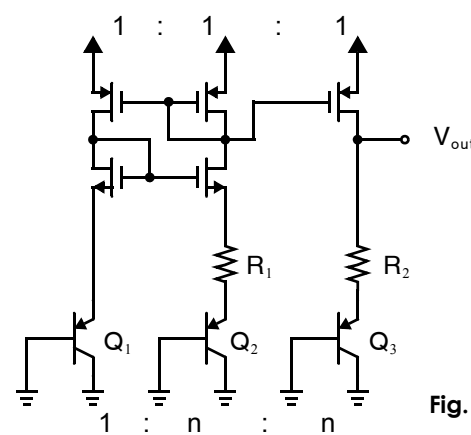


Fig. P7.17

7.7.4 Section 7.4: Voltage Regulation

7.18 A linear regulator operates from a 2.5-V supply and is used to deliver 70 mA at 1.3 V. What is the efficiency of the regulator, neglecting the power consumed in the amplifier? How much current can the amplifier consume while maintaining an efficiency better than 50%?

7.19 The linear regulator of Fig. 7.17 is to be used to deliver 12 mA of current. Using the 0.35-μm device parameters in Table 1.5, size Q_1 so that $V_{GS,1} = 800$ mV. You may assume that the body and source of Q_1 are shorted together so that the body effect may be neglected.

7.20 The LDO regulator of Fig. 7.19 is to be operated under a 1.8-V supply using the 0.18-μm device parameters in Table 1.5. It delivers 20 mA of output current at 1.2 V output voltage.

a. Determine the width of PMOS transistor Q_1 so that $V_{eff1} = 300$ mV.

b. With an output capacitance of $C_L = 10$ pF, estimate the output pole ω_{pL}.

c. Estimate the capacitances C_{gs1} and C_{gd1}.

d. What must the amplifier transconductance, G_{ma}, be in order to ensure the amplifier has a dc gain of 30-dB and that the output pole frequency is dominant so that $\omega_{pL} \approx \omega_{pa}/100$?

e. Estimate the dc current required to bias a differential pair with $V_{eff} = 150$ mV and a small-signal transconductance, $g_m = G_{ma}$. If this represents the majority of the amplifier's current consumption, what is the resulting efficiency of the linear regulator?

f. Sketch $|PSRR(\omega)|$.

CHAPTER

8 Bipolar Devices and Circuits

In the early electronic years, the majority of microcircuits were realized using bipolar-junction transistors (BJTs). However, in the late 1970s, microcircuits that used MOS transistors began to dominate the industry. Presently, the need for digital circuitry in nearly every modern integrated circuit has made CMOS devices and processing dominant, but bipolar devices remain important for many high-performance analog circuits. Modern bipolar–CMOS (BiCMOS) processes permit high performance bipolar devices to be integraed alongside large CMOS digital circuits, a particularly attractive option for mixed analog–digital applications. Thus, it remains important for an analog designer to become familiar with bipolar devices.

8.1 BIPOLAR-JUNCTION TRANSISTORS

> ***Key Point:*** *Bipolar devices offer advantages over CMOS devices when high-speed operation must be combined with high gain and/or high breakdown voltages. However, unlike MOSFETs, they draw a nonzero dc base current.*

Bipolar devices offer some inherent advantages over CMOS devices, particularly when a combination of high-frequency operation, high-gain, and/or high breakdown voltage are required. For example, in order to operate at very high speeds, modern CMOS technologies employ very thin gate dielectrics which breakdown under voltages exceeding 1–2 Volts, whereas bipolar devices can be engineered to combine high-speed operation and relatively high breakdown voltages, making them naturally suited to high-voltage radio-frequency circuits. Moreover, since BJTs can tolerate larger voltage signal swings than short gate-length MOS transistors, they are also capable of higher dynamic range. Very short gate length MOS transistors also suffer from low intrinsic gain, whereas bipolar devices suffer from no such tradeoff. Unfortunately, in bipolar transistors, the *base* control terminal has a nonzero input current when the transistor is conducting current (from the collector to the emitter for an npn transistor; from the emitter to the collector for a pnp transistor). Fortunately, at low frequencies, the base current is much smaller than the collector-to-emitter current—it may be only 0.1–1 % of the collector current for an npn transistor. For lateral pnp transistors, the base current may be as large as 1/20 of the emitter-to-collector current.

A typical cross section of an npn bipolar-junction transistor is shown in Fig. 8.1. Although this structure looks quite complicated, it corresponds approximately to the equivalent structure shown in Fig. 8.2. In a good BJT transistor, the width of the base region, W, is small (typically, less than 1 μm). Also, as we will see, the base must be more lightly doped than the emitter.

The circuit symbols used to represent npn and pnp transistors are shown in Fig. 8.3.

8.1.1 Basic Operation

To understand the operation of bipolar transistors, we consider here an npn transistor with the emitter connected to ground, as shown in Fig. 8.4. If the base voltage, V_B, is less than about 0.5 V, the transistor will be cut off, and

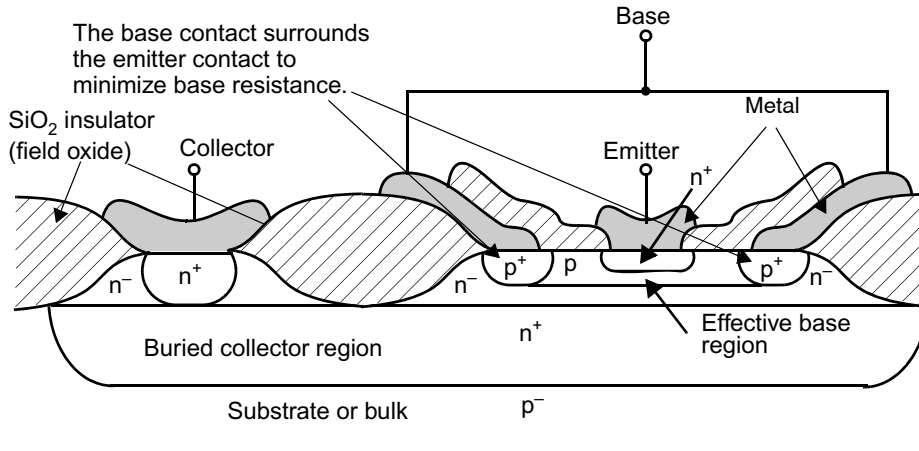


Fig. 8.1 A cross section of an npn bipolar-junction transistor.

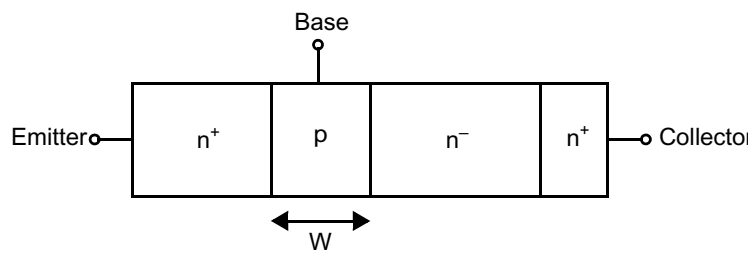


Fig. 8.2 A simplified structure of an **npn** transistor.

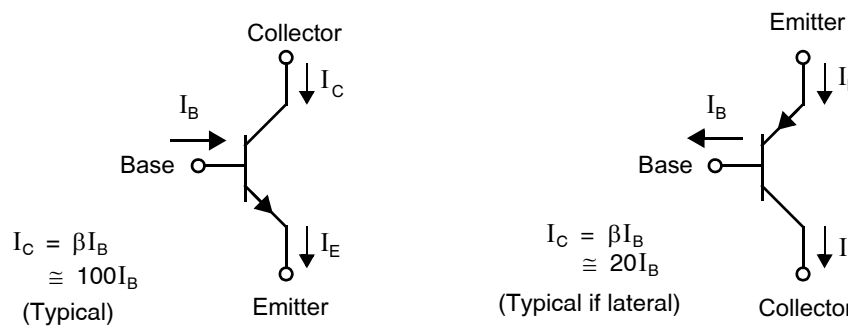


Fig. 8.3 The symbols representing (*a*) an npn bipolar-junction transistor and (*b*) a **pnp** bipolar-junction transistor.

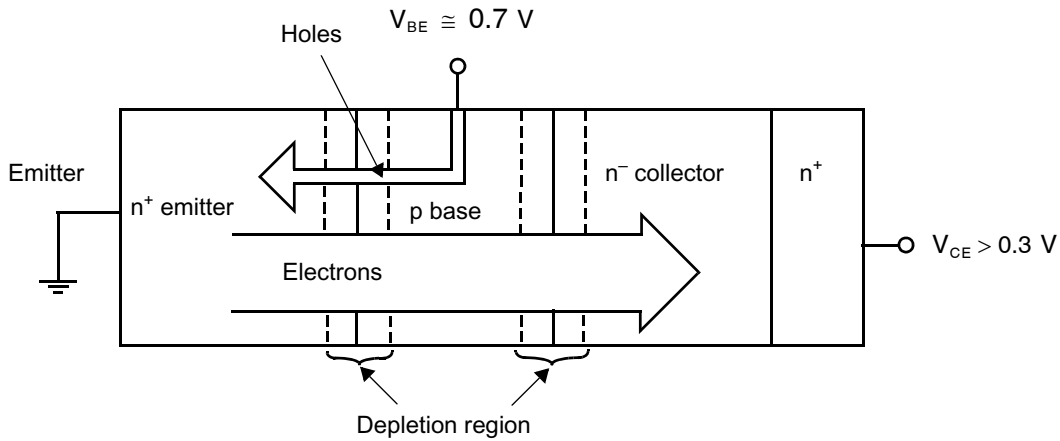


Fig. 8.4 Various components of the currents of an npn transistor.

no current will flow. We will see that when the base-emitter pn junction becomes forward biased, current will start to flow from the base to the emitter, but, partly because the base width is small, a much larger proportional current will flow from the collector to the emitter. Thus, the npn transistor can be considered a current amplifier at low frequencies. In other words, if the transistor is not cut off and the collector-base junction is reverse biased, a small base current controls a much larger collector-emitter current.

A simplified overview of how an npn transistor operates follows. When the base-emitter junction becomes forward biased, it starts to conduct, similar to any forward- biased junction. The current consists of majority carriers from the base (in this case, holes) and majority carriers from the emitter (in this case, electrons) diffusing across the junction. Because the emitter is more heavily doped than the base, there are many more electrons injected from the emitter than there are holes injected from the base. Assuming the collector voltage is large enough so that the collector-base junction is reverse biased, no holes from the base will go to the collector. However, the electrons that travel from the emitter to the base, where they are now minority carriers, diffuse away from the base-emitter junction because of the minority-carrier concentration gradient in the base region. Any of these minority electrons that get close to the collector-base junction will immediately be "whisked" across the junction due to the large positive voltage on the collector, which attracts the negatively charged electrons. In a properly designed bipolar transistor, such as that shown in Fig. 8.1, the vertical base width, W, is small, and almost all of the electrons that diffuse from the emitter to the base reach the collector-base junction and are swept across the junction, thus contributing to current flow in the collector. The result is that the collector current very closely equals the electron current flowing from the emitter to the base. The much smaller base current very closely equals the current due to the holes that flow from the base to the emitter. The total emitter current is the sum of the electron collector current and the hole base current, but since the hole current is much smaller than the electron current, the emitter current is approximately equal to the collector current.

Since the collector current is approximately equal to the electron current flowing from the emitter to the base, and the amount of this electron current is determined by the base-emitter voltage, it can be shown (see Appendix at the end of this chapter) that the collector current is exponentially related to the base-emitter voltage by the relationship

$$I_C \cong I_{CS} e^{V_{BE}/V_T} \tag{8.1}$$

where I_{CS} is the *scale current*. This scale current is proportional to the area of the base-emitter junction. The base current, determined by the hole current flowing from the base to the emitter, is also exponentially related to the

base-emitter voltage, resulting in the ratio of the collector current to the base current being a constant that, to a first-order approximation, is independent of voltage and current. This ratio, typically denoted β, is defined to be

$$\beta \equiv \frac{I_C}{I_B} \tag{8.2}$$

where I_C and I_B are the collector and base currents. Typical values of β are between 50 and 200. Lower values arise in lateral bipolar transistors, which are not optimized for high current gain, and higher values arise in hetero-junction bipolar transistors.

Key Point: *In active mode, with base-emitter junction forward-biased and base-collector junction reverse-biased, collector current increases exponentially with base-emitter voltage, and the base current is a factor β smaller. The variation in collector current with collector voltage is first-order modeled by the Early voltage.*

Note that (8.1) implies that the collector current is independent of the collector voltage. This independence ignores second-order effects such as the decrease in effective base width, W, due to the increase in the width of the collector-base depletion region when the collector bias voltage is increased. To illustrate this point, a typical plot of the collector current, I_C, as a function of collector-to-emitter voltage, V_{CE}, for different values of I_B is shown in Fig. 8.5 for a practical transistor. The fact that the curves are not flat for $V_{CE} > V_{CE\text{-sat}}$ indicates the dependence of I_C on V_{CE}. Indeed, to a good approximation, the dependence is linear with a slope that intercepts the V_{CE} axis at $V_{CE} = -V_A$ for all values of I_B. The intercept voltage value, V_A, is called the *Early voltage* for bipolar transistors, with a typical value being from 50 V to 100 V. This dependency results in a finite output impedance (as in a MOS transistor) and can be modelled by modifying equation (8.1) [Sze, 1981] to be

$$I_C \cong I_{CS} e^{V_{BE}/V_T}\left(1 + \frac{V_{CE}}{V_A}\right) \tag{8.3}$$

Large-Signal Modelling

A conducting BJT that has a V_{CE} greater than $V_{CE\text{-sat}}$ (which is approximately 0.3 V) is said to be operating in the active region. Such a collector-emitter voltage is required to ensure that none of the holes from the base go to the collector. A large-signal model of a BJT operating in the active region is shown in Fig. 8.6.

Since $I_B = I_C/\beta$, we have

$$I_B = \frac{I_{CS}}{\beta} e^{V_{BE}/V_T} = I_{BS} e^{V_{BE}/V_T} \tag{8.4}$$

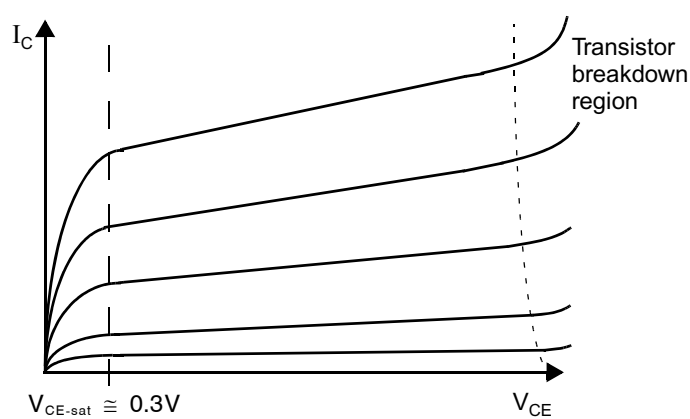


Fig. 8.5 Typical plot of I_C versus V_{CE} for a BJT.

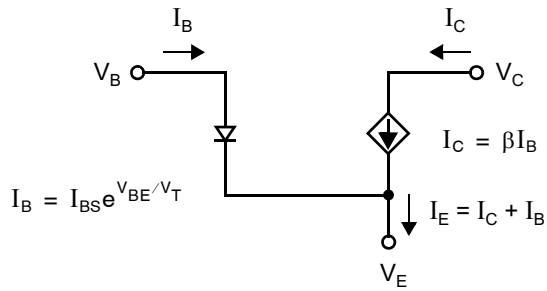


Fig. 8.6 A large-signal model for a BJT in the active region.

which is similar to a diode equation, but with a multiplying constant of $I_{CS}/\beta = I_{BS}$. Since $I_E = I_B + I_C$, we have

$$I_E = I_{CS}\left(\frac{\beta + 1}{\beta}\right) e^{V_{BE}/V_T} = I_{ES} e^{V_{BE}/V_T} \tag{8.5}$$

or equivalently

$$I_C = \alpha I_E \tag{8.6}$$

where α has been defined as

$$\alpha = \frac{\beta}{\beta + 1} \tag{8.7}$$

and for large values of β, can be approximated as

$$\alpha \cong 1 - \frac{1}{\beta} \cong 1 \tag{8.8}$$

If the effect of V_{CE} on I_C is included in the model, the current-controlled source, βI_B, should be replaced by a current source given by

$$I_C = \beta I_B\left(1 + \frac{V_{CE}}{V_A}\right) \tag{8.9}$$

where V_A is the Early-voltage constant. This additional modeling of the finite output impedance is normally not done in large-signal analysis without the use of a computer due to its complexity.

As the collector-emitter voltage approaches $V_{CE\text{-}sat}$ (typically around 0.2 to 0.3 V), the base-collector junction becomes forward biased, and holes from the base will begin to diffuse to the collector. A common model for this case, when the transistor is *saturated* or in the *saturation region*, is shown in Fig. 8.7. It should be noted that the value of $V_{CE\text{-}sat}$ decreases for smaller values of collector current.

Key Point: *As the base-collector junction approaches forward-bias, at collector-emitter voltages of about 0.2 to 0.3 V, a finite base-collector current flows, and the transistor is saturated.*

Base-Charge Storage in the Active Region

When a transistor is in the active region, many minority carriers are stored in the base region (electrons are stored in an npn transistor). Recall that this minority charge is responsible for I_C, so this charge must be removed (through the base contact) before a transistor can turn off. As in a forward-bias diode, this charge can be modelled as a diffusion capacitance, C_d, between the base and emitter given by (see Appendix at the end of this chapter)

$$C_d = \tau_b \frac{I_C}{V_T} \tag{8.10}$$

where τ_b is the base-transit-time constant. Thus, we see that the diffusion capacitance is proportional to I_C. The total base-emitter capacitance, C_{be}, will include the base-emitter depletion capacitance, C_j, in parallel with C_d. Normally, however, C_j is much less than C_d, unless the transistor current is small, and can often be ignored.

Base-Charge Storage of a Saturated Transistor

When a transistor becomes saturated, the minority-charge storage in the base and, even more so, in the lightly doped region of the collector, increases drastically. The major component of this charge storage is due to holes diffusing from the base, through the collector junction, and continuing on through the lightly doped n^- *epitaxial region* of the collector to the n^+ collector region. The n^- epitaxial region is so named because it is epitaxially grown on a p region. Most of the charge storage occurs in this region. Also, additional charge storage occurs because electrons that diffused from the collector are stored in the base, but this charge is normally smaller. The magnitude of the additional charge stored by a transistor that is saturated is given by

$$Q_s = \tau_s\left(I_B - \frac{I_C}{\beta}\right) \tag{8.11}$$

where the base overdrive current, defined to be $I_B - I_C/\beta$, is approximately equal to the hole current from the base to the collector. Normally, in saturation, $I_B >> I_C/\beta$, and (8.11) can be approximated by

$$Q_s \cong \tau_s I_B \tag{8.12}$$

The constant τ_s is approximately equal to the epitaxial-region transit time, τ_E (ignoring the storage of electrons in the base that diffused from the collector). Since the epitaxial region is much wider than the base, the constant τ_s is normally much larger than the base transit time, the constant τ_b, often by up to two orders of magnitude. The specific value of τ_s is usually found empirically for a given technology.

When a saturated transistor is being turned off, first the base current will reverse. *However, before the collector current will change, the saturation charge, Q_s, must be removed.* After Q_s is removed, the base minority charge, Q_b, will be removed. During this time, the collector current will decrease until the transistor shuts off. Typically, the time to remove Q_s greatly dominates the overall charge removal.

If the time required to remove the base saturation charge, t_s, is much shorter than the epitaxial-region transit time, τ_E, then one can derive a simple expression for the time required to remove the saturation charge. If the reverse base current (when the saturation charge is being removed), denoted by I_{BR}, remains constant while Q_s is being removed, then we have [Hodges, 1988]

$$t_s \cong \frac{Q_s}{I_{BR}} \cong \frac{\tau_s\left(I_B - \frac{I_C}{\beta}\right)}{I_{BR}} \cong \tau_s \frac{I_B}{I_{BR}} \tag{8.13}$$

where $\tau_s \cong \tau_E$.

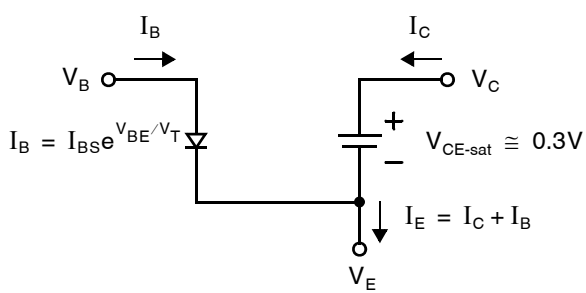


Fig. 8.7 A large-signal model for a BJT in the saturation region.

Normally, the forward base current during saturation, I_B, will be much smaller than the reverse base current during saturation-charge removal, I_{BR}. If this were not the case, then our original assumption that $t_s << \tau_E \cong \tau_s$ would not be true. In this case, the turn-off time of the BJT would be so slow as to make the circuit unusable in most applications. Nevertheless, the turn-off time for this case, when t_s is not much less than τ_E, is given by [Hodges, 1988]

$$t_s = \tau_s \ln\left[\frac{I_{BR} + I_B}{I_{BR} + \frac{I_C}{\beta}}\right] \tag{8.14}$$

The reader should verify that for $I_{BR} >> I_B$ and $I_{BR} >> I_C/\beta$, the expression in (8.14) is approximately equivalent to the much simpler one in (8.13).

In both of the cases just described, the time required to remove the storage charge of a saturated transistor is much larger than the time required to turn off a transistor in the active region. *In high-speed microcircuit designs, one never allows bipolar transistors to saturate, to avoid the long turn-off time that would result.*

Key Point: *The time required to remove the charge stored in a saturated transistor is much larger than the time required to turn off a transistor in the active region. Hence, in high-speed microcircuit designs, one never allows bipolar transistors to saturate.*

EXAMPLE 8.1

For $\tau_b = 0.2$ ns, $\tau_s = 100$ ns (a small value for τ_s), $I_B = 0.2$ mA, $I_C = 1$ mA, $\beta = 100$, and $I_{BR} = 1$ mA, calculate the time required to remove the base saturation charge using (8.13), and compare it to the time obtained using the more accurate expression of (8.14). Compare these results to the time required to remove the base minority charge for the same I_{BR}.

Solution

Using (8.13), we have

$$t_s = \frac{10^{-7}(2 \times 10^{-4})}{10^{-3}} = 20 \text{ ns} \tag{8.15}$$

Using (8.14), we have

$$t_s = 10^{-7} \ln\left[\frac{10^{-3} + 2 \times 10^{-4}}{10^{-3} + \frac{10^{-3}}{100}}\right] = 17.2 \text{ ns} \tag{8.16}$$

which is fairly close to the first result.

The time required for an active transistor to remove the base minority charge, Q_b, is given by

$$t_A = \frac{Q_b}{I_{BR}} = \frac{\tau_b I_C}{I_{BR}} = 0.2 \text{ ns} \tag{8.17}$$

This is approximately 100 times shorter than the time for removing the base saturation charge!

Small-Signal Modelling

The most commonly used small-signal model is the hybrid-π model. This model is similar to the small-signal model used for MOS transistors, except it includes a finite base-emitter impedance, r_π, and it has no emitter-to-bulk capacitance. The hybrid-π model is shown in Fig. 8.8. As in the MOS case, we will first discuss the transconductance, g_m, and the small-signal resistances, and then we will discuss the parasitic capacitances.

The transistor transconductance, g_m, is perhaps the most important parameter of the small-signal model. The transconductance is the ratio of the small-signal collector current, i_c, to the small-signal base-emitter voltage, v_{be}. Thus, we have

$$g_m = \frac{i_c}{v_{be}} = \frac{\partial I_C}{\partial V_{BE}} \tag{8.18}$$

Recall that in the active region

$$I_C = I_{CS} e^{V_{BE}/V_T} \tag{8.19}$$

Then

$$g_m = \frac{\partial I_C}{\partial V_{BE}} = \frac{I_{CS}}{V_T} e^{V_{BE}/V_T} \tag{8.20}$$

Using (8.19) again, we obtain

$$g_m = \frac{I_C}{V_T} \tag{8.21}$$

where V_T is given by

$$V_T = \frac{kT}{q} \tag{8.22}$$

Key Point: *BJT small-signal transconductance is proportional to collector current.*

and is approximately 26 mV at a room temperature of $T = 300$ K. Thus, the transconductance is proportional to the bias current of a BJT. In integrated-circuit design, it is important that the transconductance (which determines the bandwidth of many circuits) remains temperature independent, so the bias currents are usually made proportional to absolute temperature (since V_T is proportional to absolute temperature).

The presence of the resistor r_π reflects the fact that the base current is nonzero. We have

$$r_\pi = \frac{\partial V_{BE}}{\partial I_B} \tag{8.23}$$

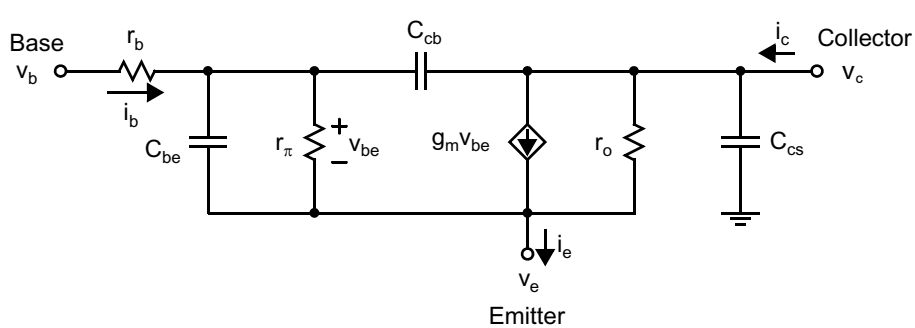


Fig. 8.8 The small-signal model of an active BJT.

Because from (8.4) we have

$$I_B = \frac{I_C}{\beta} = \frac{I_{CS}}{\beta} e^{V_{BE}/V_T} \tag{8.24}$$

we therefore have

$$\frac{1}{r_\pi} = \frac{\partial I_B}{\partial V_{BE}} = \frac{I_{CS}}{\beta V_T} e^{V_{BE}/V_T} \tag{8.25}$$

Using (8.24) again, we have

$$r_\pi = \frac{V_T}{I_B} \tag{8.26}$$

or equivalently,

$$r_\pi = \beta \frac{V_T}{I_C} = \frac{\beta}{g_m} \tag{8.27}$$

Since

$$i_e = i_c + i_b \tag{8.28}$$

we also have

$$\begin{aligned} \frac{\partial I_E}{\partial V_{BE}} &= \frac{\partial I_c}{\partial V_{BE}} + \frac{\partial I_B}{\partial V_{BE}} \\ &= g_m + \frac{g_m}{\beta} \\ &= g_m\left(\frac{1+\beta}{\beta}\right) \\ &= \frac{g_m}{\alpha} \end{aligned} \tag{8.29}$$

Continuing, we have

$$\frac{1}{r_o} = \frac{\partial I_C}{\partial V_{CE}} \tag{8.30}$$

The small-signal resistance, r_o, models the dependence of the collector current on the collector-emitter voltage. Repeating (8.3) here for convenience,

$$I_C = I_{CS} e^{V_{BE}/V_T}\left(1 + \frac{V_{CE}}{V_A}\right) \tag{8.31}$$

we have

$$\frac{1}{r_o} = \frac{\partial I_C}{\partial V_{CE}} = \frac{I_{CS}}{V_A} e^{V_{BE}/V_T} \tag{8.32}$$

Thus,

$$r_o = \frac{V_A}{I_C} \tag{8.33}$$

which is inversely proportional to the collector current.

The resistor r_b models the resistance of the semiconductor material between the base contact and the effective base region due to the moderately lightly doped base p material (see Fig. 8.1). This resistor, although small

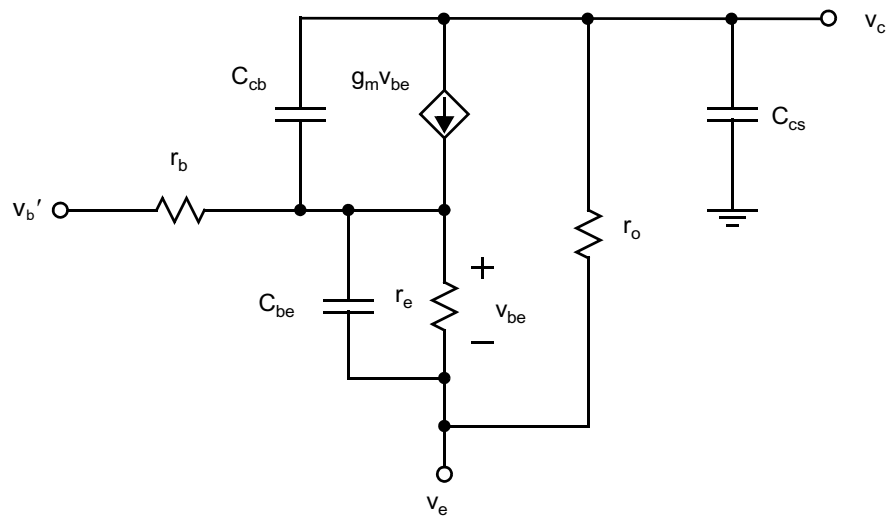


Fig. 8.9 A high-frequency, small-signal T model for an active BJT.

(typically a few hundred ohms or less), can be important in limiting the speed of very-high-frequency low-gain BJT circuits and is a major source of noise.

The hybrid-π model is only one of a number of small-signal models that can be used. One common alternative is the T model shown in Fig. 8.9. It incorporates an emitter resistance, r_e, where

$$r_e = \frac{\partial V_{BE}}{\partial I_E} = \frac{\alpha}{g_m} \tag{8.34}$$

Use of this T model can result in simplified analysis when the emitter is not at small-signal ground.

EXAMPLE 8.2

For $I_C = 1$ mA, $\beta = 100$, and $V_A = 100$ V, calculate g_m, r_π, r_e, r_o, and $g_m r_o$.

Solution

We have

$$g_m = \frac{I_C}{V_T} = \frac{10 \times 10^{-3}\ \text{A}}{0.026\ \text{V}} = 38.5\ \text{mA/V} \tag{8.35}$$

$$r_\pi = \frac{\beta}{g_m} = 2.6\ \text{k}\Omega \tag{8.36}$$

$$r_e = \frac{\alpha}{g_m} = \left(\frac{100}{101}\right)26 = 25.7\ \Omega \tag{8.37}$$

$$r_o = \frac{V_A}{I_C} = \frac{100}{10^{-3}} = 100\ \text{k}\Omega \tag{8.38}$$

The high-frequency operation of a BJT is limited by the capacitances of the small-signal model. We have already encountered one of these capacitances in Section 8.1.1: that of the forward-biased base-emitter junction, C_{be}. Recapping, we have

$$C_{be} = C_j + C_d \tag{8.39}$$

where C_j is the depletion capacitance of the base-emitter junction. For a forward-biased junction, a rough approximation for C_j is

$$C_j \cong 2A_E C_{je0} \tag{8.40}$$

The diffusion capacitance, C_d, is given in (8.10) as

$$C_d = \tau_b \frac{I_C}{V_T} = g_m \tau_b \tag{8.41}$$

Detailed expressions for τ_b are deried in the Appendix to this chapter. The capacitor, C_{cb}, models the depletion capacitance of the collector-base junction. Since this is a graded junction, we can approximate C_{cb} by

$$C_{cb} = \frac{A_C C_{jc0}}{\left(1 + \dfrac{V_{CB}}{\Phi_{c0}}\right)^{1/3}} \tag{8.42}$$

where A_C is the effective area of the collector-base interface.

Due to the lower doping levels in the base and especially in the collector (perhaps 5×10^{22} acceptors/m^3 and 10^{21} donors/m^3, respectively), Φ_{c0}, the built-in potential for the collector-base junction, will be less than that of the base-emitter junction (perhaps 0.75 V as opposed to 0.9 V). It should be noted that the cross-sectional area of the collector-base junction, A_C, is typically much larger than the effective area of the base-emitter junction, A_E, which is shown in Fig. 8.1. This size differential results in $A_C C_{jc0}$ being larger than $A_E C_{je0}$, the base-emitter junction capacitance at 0 V bias, despite the lower doping levels.

> ***Key Point:*** *The high-frequency operation of active BJTs is limited by capacitances modeling charge storage in the depletion regions and the storage of diffusing charge-carriers in the base.*

Finally, another large capacitor is C_{cs}, the capacitance of the collector-to-substrate junction. Since this area is quite large, C_{cs}, which is the depletion capacitance that results from this area, will be much larger than either C_{cb} or the depletion capacitance component of C_{be}, that is, C_j. The value of C_{cs} can be calculated using

$$C_{cs} = \frac{A_T C_{js0}}{\left(1 + \dfrac{V_{CS}}{\Phi_{s0}}\right)^{1/2}} \tag{8.43}$$

where A_T is the effective transistor area and C_{js0} is the collector-to-substrate capacitance per unit area at 0-V bias voltage.

8.1.2 Analog Figures of Merit

The intrinsic gain of a bipolar transistor, A_i, as with MOS transistors, is the maximum possible gain achievable using that single transistor and is given by

$$A_i = g_m r_o = V_A / V_T \tag{8.44}$$

Key Point: *Active BJTs offer an intrinsic gain that is generally higher than that of MOS transistors and independent of operating point.*

Hence, the intrinsic gain of a bipolar transistor is a constant value independent of the transistor operating point. This is fundamentally different than a MOS transistor, whose intrinsic gain was shown in equation (1.115) $A_i \cong 2/\lambda V_{eff}$ to decrease with increasing bias current. The intrinsic gain of a npn BJT is usually between 2000 and 8000.

EXAMPLE 8.3

Find the maximum possible gain of a single-transistor amplifier using the transistor in Example 8.2.

Solution

The maximum gain is given by $g_m r_o$,

$$g_m r_o = \frac{V_A}{V_T} = 3{,}846 \tag{8.45}$$

Note that this gain is much higher than the 32.9 that was found for a single MOS transistor in Example 1.15. Also note that this BJT maximum gain is independent of the bias current. For MOS transistors, equation (1.115) shows that the maximum gain is inversely proportional to V_{eff} and, hence, inversely proportional to the square-root of its bias current. This is one of the reasons why it is possible to realize a single-transistor BJT amplifier with a much larger gain than would result if a MOS transistor were used, especially at high current levels (and therefore at high frequencies).

A common indicator for the *speed* of a BJT is the frequency at which the transistor's current gain drops to unity, when its collector is connected to a small-signal ground. This frequency is denoted f_t and is called the transistor *unity-gain frequency* or *cutoff frequency.* We can see how this frequency is related to the transistor model parameters by analyzing the small-signal circuit of Fig. 8.10. In the simplified model in Fig. 8.10(*b*), the resistor

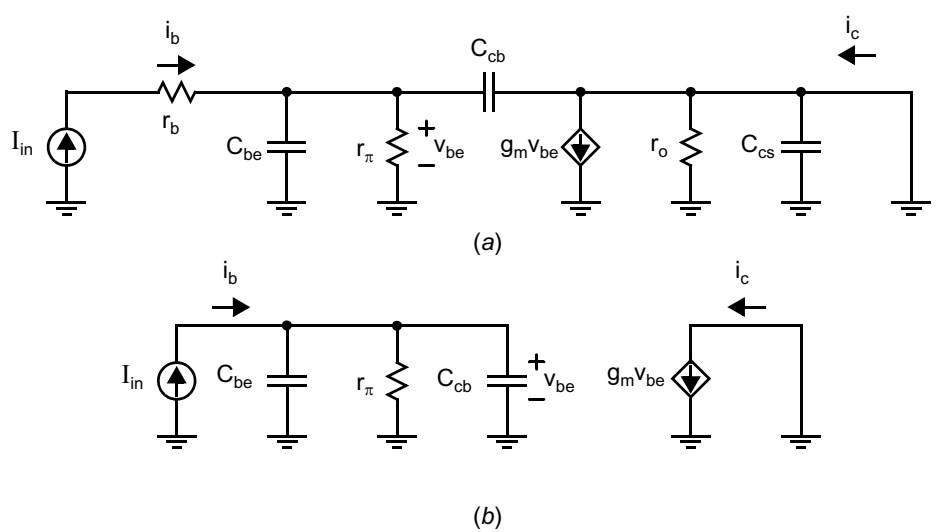


Fig. 8.10 (*a*) A small-signal model used to find f_t; (*b*) an equivalent simplified model.

r_b is ignored because it has no effect on i_b since the circuit is being driven by a perfect current source. We have

$$V_{be} = i_b\left(r_\pi \left\| \frac{1}{sC_{be}} \right\| \frac{1}{sC_{cb}}\right) \tag{8.46}$$

and

$$i_c = g_m V_{be} \tag{8.47}$$

Solving for i_c / i_b gives

$$\frac{i_c}{i_b} = \frac{g_m r_\pi}{1 + s(C_{be} + C_{cb})r_\pi} \tag{8.48}$$

At low frequencies, the current gain is $g_m r_\pi$, which equals the expected value of β using (8.27). At high frequencies, i_c / i_b is approximately given by

$$\left|\frac{i_c}{i_b}(\omega)\right| \cong \frac{g_m r_\pi}{\omega(C_{be} + C_{cb})r_\pi} = \frac{g_m}{\omega(C_{be} + C_{cb})} \tag{8.49}$$

To find the unity-gain frequency, we set $\left|(i_c/i_b)(\omega_t)\right| = 1$ and solve for ω_t, which results in

$$\omega_t = \frac{g_m}{C_{be} + C_{cb}} \tag{8.50}$$

or

$$f_t = \frac{g_m}{2\pi(C_{be} + C_{cb})} \tag{8.51}$$

Substituting (8.21) and (8.39) into (8.51) yields

$$f_t = \left[2\pi\left(\frac{(C_{be} + C_{cb})V_T}{I_C} + \tau_b\right)\right]^{-1} \tag{8.52}$$

Finally, using the expression (8.102) for the transit time of the base from the Appendix gives

$$f_t = \left[2\pi\left(\frac{(C_{be} + C_{cb})V_T}{I_C} + \frac{W^2}{2D_n}\right)\right]^{-1} \tag{8.53}$$

This expression neglects some other charge storage effects, but captures the most important terms. At low collector currents, the unity-gain frequency is dominated by the terms inversely proportional to I_C. As collector current increases, those terms diminish and f_t increases up to a limit imposed by the base transit time constant, τ_b

$$f_{t,\max} \approx \frac{1}{2\pi} \cdot \frac{2D_n}{W^2} \tag{8.54}$$

where D_n is the diffusion constant of minority carriers in the base, $(\mu_n kT)/q$, and W is the width of the base region. Clearly, the width of the base region is a critical dimension influencing the high-speed performance of a bipolar transistor, analogous to the length of a MOSFET gate. Often, either f_t, ω_t, or $\tau_t = 1/\omega_t$ will be specified for a transistor at a particular bias current. These values are representative of an upper limit on the maximum frequency at which the transistor can be effectively used at that bias current.

> ***Key Point:*** *BJT unity-gain frequency increases in proportion to bias current up to a limit imposed by the transit time of carriers through the base. Hence, the base thickness is a critical parameter limiting the maximum operating speed of a BJT.*

Not captured by (8.53) are speed-limiting effects that arise at high current densities. For example, at high current densities the concentration of charge carriers passing through the base becomes so high that it moves the effective edge of the base region further towards and even into the collector. Doing so increases the effective width of the base, W, resulting in an

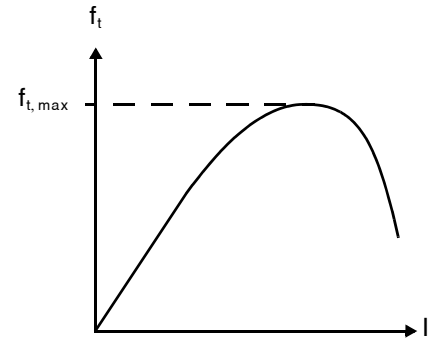


Fig. 8.11 A sketch of transistor f_t versus collector current, I_C.

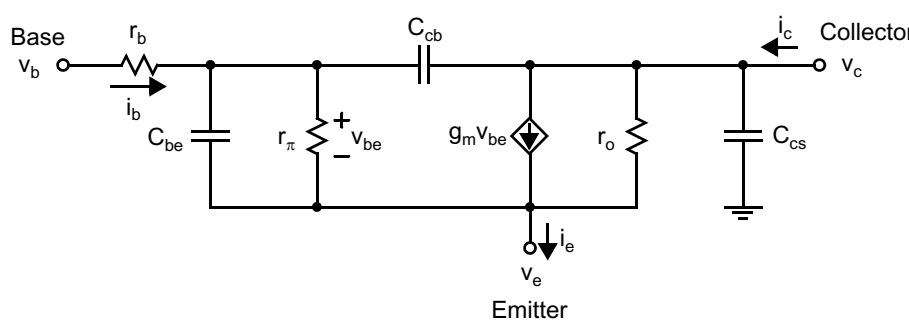
increase in the transit time of the base and a decrease in f_t. This is known as the Kirk effect [Kirk, 1962] and causes f_t to drop at high collector currents as illustrated in Fig. 8.11.

8.2 BIPOLAR DEVICE MODEL SUMMARY

Active Transistor

$I_C = I_{CS} e^{V_{BE}/V_T}$	$V_T = \frac{kT}{q} \cong 26$ mV at 300 °K
For more accuracy, $I_C = I_{CS} e^{V_{BE}/V_T}\left(1 + \frac{V_{CE}}{V_A}\right)$	
$I_{CS} = \frac{A_E q D_n n_i^2}{W N_A}$	$I_B = \frac{I_C}{\beta}$
$I_E = \left(1 + \frac{1}{\beta}\right) I_C = \frac{1}{\alpha} I_C = (\beta + 1) I_B$	$\beta = \frac{I_C}{I_B} = \frac{D_n N_D L_p}{D_p N_A W} \cong 2.5 \frac{N_D L_p}{N_A W}$
$\alpha = \frac{\beta}{1 + \beta}$	

Small-Signal Model of an Active BJT



$g_m = \frac{I_C}{V_T}$	$r_\pi = \frac{V_T}{I_B} = \frac{\beta}{g_m}$
$r_e = \frac{\alpha}{g_m}$	$r_o = \frac{V_A}{I_C}$
$g_m r_o = \frac{V_A}{V_T}$	$C_{be} = C_j + C_d$
$C_j \cong 2A_E C_{je0}$	$C_d = \tau_b \frac{I_C}{V_T} = g_m \tau_b$
$C_{cb} = \frac{A_C C_{jc0}}{\left(1 + \frac{V_{CB}}{\Phi_{c0}}\right)^{1/3}}$	$C_{cs} = \frac{A_T C_{js0}}{\left(1 + \frac{V_{CS}}{\Phi_{s0}}\right)^{1/2}}$

8.3 SPICE MODELING

For historical reasons, most parameters for modelling bipolar transistors are specified absolutely. Also, rather than specifying the emitter area of a BJT in $(\mu m)^2$ on the line where the individual transistor connections are specified, most SPICE versions have multiplication factors. These multiplication factors can be used to automatically multiply parameters when a transistor is composed of several transistors connected in parallel. This multiplying parameter is normally called M.

The most important dc parameters are the transistor current gain, β, specified by the SPICE parameter BF; the transistor-transport saturation current, I_{CS}, specified using the parameter IS; and the Early-voltage constant, specified by the parameter VAF. Typical values for these might be 100, 10^{-17} A, and 50 V, respectively. If one wants to model the transistor in reverse mode (where the emitter voltage is higher than the collector voltage for an npn), then one might specify BR, ISS, and VAR, as well; these are the parameters that correspond to BF, IS, and VAF in the reverse direction. Typically, this reverse-mode modelling is not important for most circuits. Some other important dc parameters for accurate simulations are the base, emitter, and collector resistances, which are specified by RB, RE, and RC, respectively. It is especially important to specify RB (which might be 200 Ω to 500 Ω).

The important capacitance parameters and their corresponding SPICE parameters include the depletion capacitances at 0-V bias voltage, CJE, CJC, CJS; their grading coefficients, MJE, MJC, MJS; and their built-in voltages, VJE, VJC, VJS, for base-emitter, base-collector, and collector-substrate junctions. Again, the 0-V depletion capacitances should be specified in absolute values for a unit-sized transistor. Normally the base-emitter and base-collector junctions are graded (i.e., MJE, MJC = 0.33), whereas the collector-substrate junction may be either abrupt (MJS = 0.5) or graded (MJS = 0.33), depending on processing details. Typical built-in voltages might be 0.75 V to 0.8 V. In addition, for accurate simulations, one should specify the forward-base transit time, τ_F, specified by TF, and, if the transistor is to be operated in reverse mode or under saturated conditions, the reverse-base transit time, τ_R, specified by TR.

The most important of the model parameters just described are summarized in Table 8.1. These form the basis of the Gummel-Poon SPICE model, and its variants. Many other parameters can be specified if accurate simulation is desired. Other parameters might include those to model β degradation under very high or low current applications and parameters for accurate noise and temperature analysis.

Table 8.1 The most important SPICE parameters for modelling BJTs.

SPICE Parameter	Model Constant	Brief Description	Typical Value
BF	b	Transistor current gain in forward direction	100
BR	β_R	Transistor current gain in the reverse direction	1
IS	I_{CS}	Transport saturation current in forward direction	2×10^{-18} A
VAF	V_A	Early voltage in forward direction	50 V
RB	r_b	Series base resistance	500 Ω
RE	R_E	Series emitter resistance	30 Ω
RC	R_C	Series collector resistance	90 Ω
CJE	C_{je0}	Base-emitter depletion capacitance at 0 V	0.015 pF
CJC	C_{jc0}	Base-collector depletion capacitance at 0 V	0.018 pF
CJS	C_{js0}	Collector-substrate depletion capacitance at 0 V	0.040 pF
MJE	m_e	Base-emitter junction exponent (grading factor)	0.30
MJC	m_c	Base-collector junction exponent (grading factor)	0.35
MJS	m_s	Collector-substrate junction exponent (grading factor)	0.29
VJE	Φ_e	Base-emitter built-in potential	0.9 V
VJC	Φ_c	Base-collector built-in potential	0.7 V
VJS	Φ_s	Collector-substrate built-in potential	0.64 V
TF	τ_F	Forward-base transit time	12 ps
TR	τ_R	Reverse-base transit time	4 ns

Another popular SPICE model for modern bipolar transistors is the High Current Model (HiCuM). As the name suggests, it is particularly effective for modeling transistor behavior at high current densities, and is therefore well suited to the design of circuits operating at very high frequencies. Since high-frequency design is an area where bipolar devices continue to offer some advantages over MOSFETs, the HiCuM model has become very popular.

Readers should refer to their SPICE manuals for complete descriptions of the Gummel-Poon and High-Current models and their parameters.

8.4 BIPOLAR AND BICMOS PROCESSING

The processing steps required for realizing bipolar transistors are similar to those used for realizing MOS transistors, although naturally the precise recipe is different. Thus, rather than presenting the complete realization of modern bipolar transistors, we briefly discuss some of the modifications needed for realizing them.

8.4.1 Bipolar Processing

A bipolar process normally starts with a p^- substrate. The first masking step involves the diffusion (or ion implantation) of n^+ regions into the substrate wherever transistors are desired. These n^+ regions are used to lower the series collector resistance. Next, an n^- single-crystal epitaxial layer is deposited.

In the next step, isolation structures are patterned, using either field-oxide growth and/or by etching trenches and filling them with SiO_2. Next, the n^+-collector contact region is implanted. This region extends from the surface down to the n^+ -buried region under the transistor.

Polysilicon is generally used to contact the emitter, the base, and possibly the collector, as Fig. 8.12 shows for a typical npn BJT structure. One approach is to begin by depositing the base p^+ polysilicon. This polysilicon is heavily doped p^+ so that later, during a high-temperature step, the boron dopant from the polysilicon contact diffuses into the silicon underneath the base polysilicon to make the underlying region p^+. The base polysilicon is removed in the active area of the transistor. Next, using one of a variety of possible methods, the base polysilicon is covered with a thin layer of SiO_2, perhaps 0.5 μm in thickness, leaving an opening over the active area of the transistor. This SiO_2 *spacer,* which separates the base polysilicon from the emitter polysilicon, as shown in Fig. 8.12, allows the base polysilicon (or contact) to be very close to the emitter polysilicon (or contact), thereby minimizing the base resistance. Next, the base is ion-implanted to p -type silicon, and then n^+ polysilicon is deposited for the emitter. At this point, the true emitter has not yet been formed—only the emitter n^+ polysilicon has been laid down. However, when the wafer is annealed, the n^+ from the emitter polysilicon diffuses into the base p silicon to form the true emitter region. During annealing, the p^+ dopants from the base polysilicon also diffuse into the extrinsic base region. As a result, this procedure results in a self-aligned process since the use of the SiO_2 spacer allows the base polysilicon to determine where the emitter is finally located. The importance of this process is that, through the use of self-aligned contacts and field-oxide isolation, very small, high-frequency bipolar transistors can be realized using methods similar to those used in realizing modern MOS transistors.

8.4.2 Modern SiGe BiCMOS HBT Processing

A major limitation of the processing described above is that the p -type base region is formed by ion implantation. As illustrated in Fig. 2.4, this results in a dopant profile with an approximately Gaussian distribution. Moreover, any subsequent high-temperature processing of the wafer may cause annealing and, hence, a spreading of the base

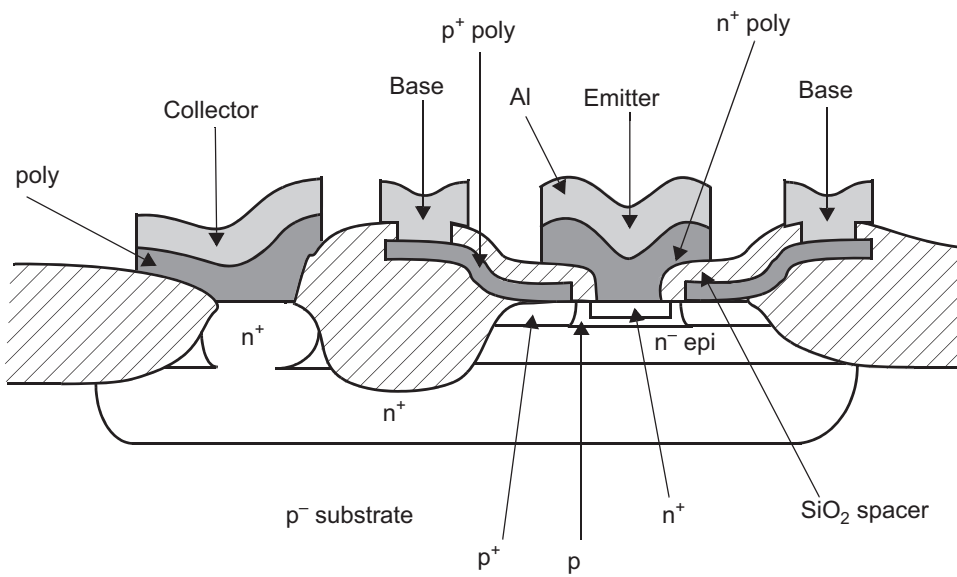


Fig. 8.12 Cross section of a modern, self-aligned bipolar transistor with oxide isolation. The term "poly" refers to polysilicon.

region. As a result, it is difficult to realize a very thin base region using this method, which as shown in equation (8.54) limits the device's high-frequency performance.

Key Point: *High-performance modern bipolar transistors are manufactured by epitaxially growing a very thin base region. The use of a different material for the base, such as silicon-germanium, results in a heterojunction bipolar transistor (HBT) whose bandgap structure can be tailored to provide high current gain, high unity-gain frequency, and/or high Early voltage.*

The solution is to epitaxially grow the base region on top of the collector. This allows very thin base regions to be deposited with very abrupt doping profiles at the junctions. It also affords the opportunity to grow a compound semiconductor for the base region instead of pure silicon. Specifically, the use of silicon-germanium (SiGe) in the base offers several advantages [Harame, 1995]. Because the base region is SiGe while the collector and emitter are Si, this device structure is a *heterojunction bipolar transistor* (HBT). Introducing germanium in the base lowers its bandgap, which in turn lowers the energy required for electrons to cross the emitter-base junction (in a npn transistor). Hence, at a fixed V_{BE}, the electron current is higher with a SiGe base layer than a silicon base layer, resulting in increased current gain β. The increase in β provided by the use of a SiGe base offsets the need to have a lower dopant concentration there; hence, the dopant concentration of the base can be higher in a SiGe HBT than in a silicon BJT, lowering the base resistance r_b. Moreover, by varying the environment in during epitaxial growth of the base, it is possible to vary the silicon-germanium ratio across the base region. Doing so gives rise to an electric field which accelerates electrons from emitter to collector, reducing the base transit time τ_b and increasing the unity-gain frequency f_t compared with silicon transistors. Grading the germanium content also makes it possible to increase the Early voltage, V_A. In summary, the use of an epitaxially grown base region is a significant advance that has offered the flexibility to engineer a bipolar transistor with improved analog performance in virtually every way.

All of the bipolar processing described above can be performed on wafers alongside CMOS circuitry. A series of masks defines the regions where n^+ subcollectors are deposited, base epitaxy is performed, emitter contacts are made, etc. without interfering with neighbouring CMOS circuits. The primary challenges are ensuring that the surface preparation required for epitaxial growth of the base does not deteriorate the CMOS regions, and that high-temperature annealing of the CMOS devices does not deteriorate bipolar transistor performance. The result is a bipolar-CMOS (BiCMOS) manufacturing processes that can realize high-performance bipolar devices alongside CMOS devices.

8.4.3 Mismatch in Bipolar Devices

Like all integrated circuit components, side-by-side bipolar transistors on the same die that are specified to have the exact same dimensions will still exhibit slight parametric differences due to finite tolerances during fabrication. Although bipolar circuits are generally less affected by mismatch than their MOS counterparts, mismatch in the collector currents of bipolar transistor is of significant interest as it leads to offset in differential pairs and current mirrors. Collector current mismatch is determined by mismatch between the transistors' scale currents.

$$I_{CS} = \frac{A_E q D_n n_i^2}{W N_A} \tag{8.55}$$

Uncertainty in the emitter area A_E, base width W, and base dopant concentration N_A, during manufacture are all causes of mismatch in bipolar transistor scale currents, and hence collector currents. As a result, collector currents vary statistically as described in Section 2.3.3 so that the percentage mismatch between the collector current of two identically sized bipolar transistors is inversely related to the device area. Specifically, the mean squared deviation in the collector current of a particular bipolar device away from its nominal value is modeled as follows:

$$\sigma^2\left(\frac{\Delta I_C}{I_C}\right) = \sigma^2\left(\frac{\Delta I_{CS}}{I_{CS}}\right) = \frac{A_{I_C}^2}{A_E} \tag{8.56}$$

The constant A_{Ic} parameterizes the matching accuracy of a particular bipolar fabrication process, and is in the range of 0.01–0.04 μm^{-1} for typical modern bipolar processes [Tuinhout, 2003].

8.5 BIPOLAR CURRENT MIRRORS AND GAIN STAGES

In this section, we look at bipolar current mirrors and two bipolar gain circuits: the emitter follower and the differential pair. Although much of the small-signal analyses follow those of their MOS counterparts very closely, one difference is the finite base impedance of bipolar transistors. Here we shall see some simple rules for dealing with this finite base impedance during small-signal analysis. Because of its importance in translinear circuits, we also look at the *large-signal* behavior of the differential pair.

8.5.1 Current Mirrors

The most popular bipolar current mirrors are very similar to the MOS current mirrors. A simple bipolar current mirror is shown in Fig. 8.13. This current mirror has an output current almost equal to its input current. In fact, its output current is slightly smaller than the input current, due to the finite base currents of Q_1 and Q_2. Taking these two base currents into account results in

$$I_{out} = \frac{1}{(1 + 2/\beta)} I_{in} \tag{8.57}$$

where β is the transistor current gain. For large β, this relation is approximately given by $I_{out} \cong (1 - 2/\beta) I_{in}$.

In one commonly used modification of the simple bipolar current mirror, an emitter-follower buffer, Q_3, is added to supply the base currents, as shown in Fig. 8.14. This additional transistor minimizes the errors due to finite base currents, resulting in $I_{out} \cong I_{in}(1 - 2/\beta^2)$. Such an arrangement is almost always used for current mirrors when lateral transistors[1] are used because of their low current gains (i.e., β's on the order of only 10–20). The output impedance of both of the previous current sources is equal to the output impedance of Q_2, which is r_{o2}.

> ***Key Point:*** *Bipolar current mirrors are similar to CMOS mirrors, except that finite base current is a source of mismatch between the input and output currents.*

In another often-used variation of the simple current mirror, emitter degeneration is added. This approach results in larger output impedances and also minimizes errors caused by mismatches between Q_1 and Q_2. An example of this current mirror is shown in Fig. 8.15. In an analysis similar to that given for the MOS current mirror with source degeneration, the output impedance is now found to be given by

$$r_{out} \cong r_{o2}(1 + g_{m2}R_e) \tag{8.58}$$

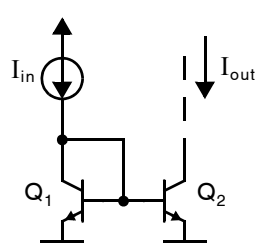


Fig. 8.13 A simple bipolar current mirror.

1. In many bipolar processes, pnp transistors are only available as lateral devices.

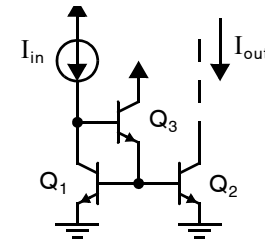


Fig. 8.14 A current mirror with fewer inaccuracies caused by finite base currents.

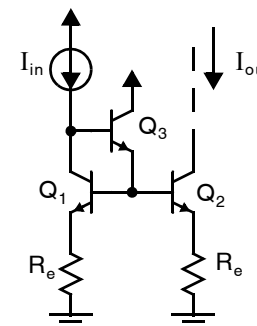


Fig. 8.15 A current mirror with emitter degeneration.

for $R_e << r_\pi$. Normally, this current mirror is designed so that the bias voltage across R_e, defined to be V_{Re}, is about 0.25 V. This implies that R_e is given by

$$R_e = \frac{V_{Re}}{I_{e2}} \cong \frac{V_{Re}}{I_{c2}} \tag{8.59}$$

Using the relationship $g_{m2} = I_{c2}/V_T$, where $V_T = kT/q \cong 26$ mV at 300 °K, gives

$$r_{out} \cong r_{o2}(1 + g_{m2}R_e) = r_{o2}\left(1 + \frac{V_{Re}}{V_T}\right) \cong 11r_{o2} \tag{8.60}$$

assuming $V_{Re} = 0.25$ V. It should be mentioned here that the addition of these R_e resistors also minimizes the noise output current of this current mirror, generated by the base resistance thermal noise, which is often the major source of noise in bipolar wideband circuits.

To achieve still higher output impedances, one can use either a cascode or a Wilson current mirror, as shown in Fig. 8.16. The Wilson current mirror is preferred in bipolar realizations because the cascode mirror exhibits large errors, due to the fact that the base currents of all of the transistors are supplied by I_{in} only. As a result, the output current is smaller than the input current by a factor roughly equal to $1–4/\beta$. For the Wilson current mirror, the base currents of Q_3 and Q_4 are supplied by I_{in}, whereas the base currents of Q_1 and Q_2 come from I_{out}. It can be shown that the errors due to finite base currents are on the order of $2/\beta^2$ [Gray, 2009]. It can also be shown that both of these current mirrors have an output impedance on the order of [Gray, 2009]

$$r_{out} \cong \frac{\beta r_o}{2} \tag{8.61}$$

8.5.2 Emitter Follower

A bipolar emitter-follower is very similar to a MOS source follower, except that its input resistance is not infinite and its gain is normally much closer to unity. The analysis of its low-frequency response provides a good

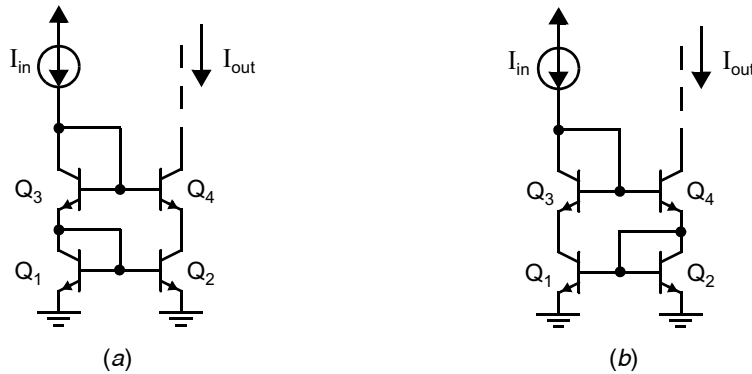


Fig. 8.16 High-output impedance (*a*) cascode and (*b*) Wilson current mirrors.

illustration of the use of the bipolar T model, and also of some relatively general principles for coming up with quick estimates of input and output impedances of bipolar circuits.

A bipolar emitter follower with a resistive load is shown in Fig. 8.17. Its small-signal model at low frequencies, where the bipolar T model has been used, is shown in Fig. 8.18(*a*). In this small-signal model, the device base resistance r_b has been ignored, since it can be easily taken into account after the fact by simply making R_S slightly larger (i.e., equal to $R_S + r_b$). Notice that R_E and r_o are in parallel. This allows a slightly simplified small-signal model to be analyzed, as shown in Fig. 8.18(*b*), where $R'_E = R_E \parallel r_o$. The analysis of the gain of this circuit is done in two steps. First, the input impedance looking into the base of the transistor, R_b, is found, allowing us to calculate the gain from the input to the base. Second, the gain from the base to the output is found, which allows us to derive the overall gain. The current in the emitter, i_e, is given by

$$i_e = i_b(\beta + 1) \tag{8.62}$$

Therefore,

$$v_b = i_e(r_e + R'_E) = i_b(\beta + 1)(r_e + R'_E) \tag{8.63}$$

This gives

$$R_b = \frac{v_b}{i_b} = (\beta + 1)(r_e + R'_E) \tag{8.64}$$

Alternatively, noting that

$$(\beta + 1)r_e = (\beta + 1)\frac{\alpha}{g_m} = (\beta + 1)\frac{\beta/(\beta + 1)}{g_m} = \frac{\beta}{g_m} = r_\pi \tag{8.65}$$

gives

$$R_b = r_\pi + (\beta + 1)\, R'_E \tag{8.66}$$

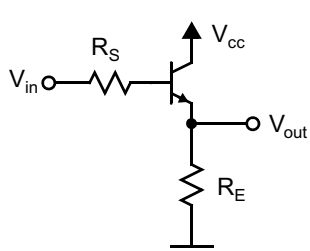


Fig. 8.17 A bipolar emitter follower.

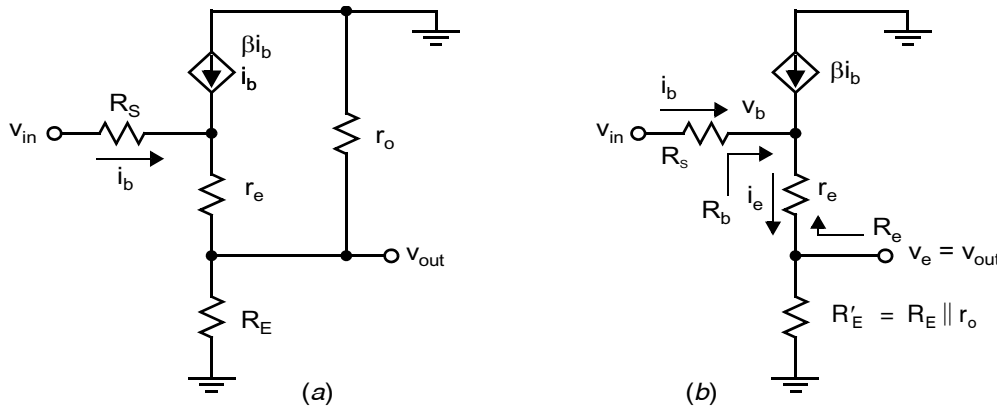


Fig. 8.18 (*a*) The small-signal model for the bipolar emitter follower and (*b*) a simplified model.

Equations (8.64) and (8.66) illustrate a principle that is generally applicable for bipolar circuits: *At low frequencies, resistances in series with the emitter appear* $\beta + 1$ *times larger when seen looking into the base or, equivalently, when they are reflected into the base.* Continuing, using the resistor-divider formula, we now have

Key Point: *At low frequencies, resistances in series with the emitter appear (β+1) times larger when seen looking into the base.*

$$\frac{v_b}{v_{in}} = \frac{R_b}{R_b + R_S} = \frac{(\beta + 1)(r_e + R'_E)}{(\beta + 1)(r_e + R'_E) + R_S} \tag{8.67}$$

The gain from the base to the emitter, which is the output, is easily found from Fig. 8.17(*b*), again using the resistor-divider formula, to be given by

$$\frac{v_{out}}{v_b} = \frac{R'_E}{R'_E + r_e} \cong \frac{R'_E}{R'_E + 1/g_m} \tag{8.68}$$

Using (8.67) and (8.68), the overall gain is now given by

$$\frac{v_{out}}{v_{in}} = \frac{v_b}{v_{in}}\frac{v_{out}}{v_b} = \left(\frac{(\beta + 1)(r_e + R'_E)}{(\beta + 1)(r_e + R'_E) + R_S}\right)\left(\frac{R'_E}{R'_E + 1/g_m}\right) \tag{8.69}$$

With a little practice, transfer functions such as that given by (8.69) can be written by simply inspecting the actual circuit, without actually analyzing the small-signal model.

It is also interesting to find the output impedance of the emitter follower at low frequencies excluding R'_E, which is equal to the impedance seen looking into the emitter. The small-signal model for this analysis is shown in Fig. 8.19, where the input source has been set to 0 and the impedance $R_e = v_x/i_x$ is to be found. First note that

$$i_b = \frac{i_e}{\beta + 1} = \frac{-i_x}{\beta + 1} \tag{8.70}$$

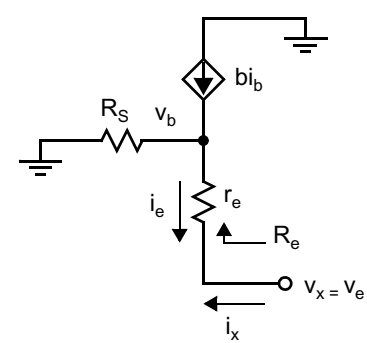


Fig. 8.19 The small-signal model for finding the output impedance of an emitter follower.

since $i_x = -i_e$. Therefore, we have

$$\begin{aligned} v_x &= v_b + i_x r_e \\ &= -i_b R_S + i_x r_e \\ &= i_x \frac{R_S}{\beta + 1} + i_x r_e \end{aligned} \tag{8.71}$$

This gives the impedance seen looking into the emitter as

$$R_e = \frac{v_x}{i_x} = \frac{R_S}{\beta + 1} + r_e = \frac{R_S}{\beta + 1} + \frac{r_\pi}{\beta + 1} \tag{8.72}$$

This is an example of a general principle for bipolar circuits: *Resistances in series with the base are divided by* $\beta + 1$ *when they are seen looking into the emitter, or, equivalently, are reflected to the emitter.* The total output impedance of the emitter-follower is now simply R_e in parallel with R_E'.

> ***Key Point:*** *Resistances in series with the base are divided by (β+1) when they are seen looking into the emitter.*

The high-frequency analysis of the emitter follower is very similar to that of the source follower presented in Section 4.4. As with source followers, some care is required to avoid overshoot and ringing in emitter followers. In fact, ringing is somewhat more likely with emitter followers than with source followers due in part to the absence of any body effect.

8.5.3 Bipolar Differential Pair

Large-Signal

A bipolar differential pair is shown in Fig. 8.20. This circuit's large-signal behavior can be analyzed by first recalling the exponential relationship for a bipolar transistor,

$$I_C = I_S e^{(V_{BE}/V_T)} \tag{8.73}$$

which can be used to find the base-emitter voltages

$$V_{BE1} = V_T \ln\left(\frac{I_{C1}}{I_{S1}}\right) \tag{8.74}$$

$$V_{BE2} = V_T \ln\left(\frac{I_{C2}}{I_{S2}}\right) \tag{8.75}$$

Now, writing an equation for the sum of voltages around the loop of input and base-emitter voltages, we have

$$V^+ - V_{BE1} + V_{BE2} - V^- = 0 \tag{8.76}$$

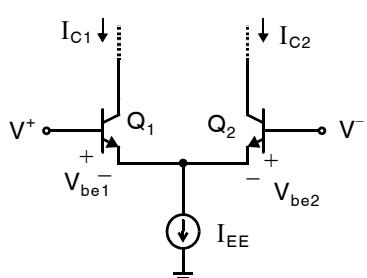


Fig. 8.20 A bipolar differential pair.

Combining (8.74), (8.75), and (8.76), and assuming $I_{S1} = I_{S2}$, we can write

$$\frac{I_{C1}}{I_{C2}} = e^{(V^+ - V^-)/V_T} = e^{V_{id}/V_T} \tag{8.77}$$

where V_{id} is defined as the difference between V^+ and V^-. In addition, we can write

$$\alpha I_{EE} = I_{C1} + I_{C2} \tag{8.78}$$

where α is defined to be $\beta/(\beta + 1)$ and is due to some currents flowing through the base terminals. Finally, combining (8.77) and (8.78), we have

$$I_{C1} = \frac{\alpha I_{EE}}{1 + e^{-V_{id}/V_T}} \tag{8.79}$$

$$I_{C2} = \frac{\alpha I_{EE}}{1 + e^{V_{id}/V_T}} \tag{8.80}$$

A plot of these two currents with respect to V_{id} is shown in Fig. 8.21, where we note that the currents are split equally at $V_{id} = 0$ and saturate near I_{EE} or 0 for differential input voltages approaching 4 V_T (around 100 mV). Note that the relations shown in (8.79) and (8.80) results in a hyperbolic tangent function when their difference is taken.

$$I_{C2} - I_{C1} = \alpha I_{EE} \tanh(V_{id}/2V_T) \tag{8.81}$$

Thus this current–voltage relationship for a bipolar differential is commonly referred to as the *tanh relationship.*

Small-Signal

The small-signal model for the bipolar differential pair of Fig. 8.20 is shown in Fig. 8.22. Once again, defining $v_{id} = v^+ - v^-$, we have

$$i_{c1} = \alpha i_{e1} = \frac{\alpha v_{id}}{r_{e1} + r_{e2}} = \frac{\alpha v_{id}}{(\alpha/g_{m1}) + (\alpha/g_{m2})} \tag{8.82}$$

In the case where both transistors have the same bias currents through them (i.e., the nominal differential voltage is 0, resulting in $g_{m1} = g_{m2}$), we find the same result as for a MOS differential pair,

$$i_{c1} = \frac{g_{m1}}{2} v_{id} \tag{8.83}$$

A similar result holds for i_{c2}.

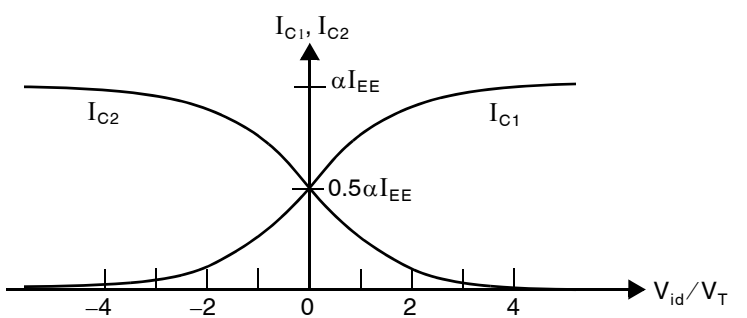


Fig. 8.21 Collector currents for a bipolar differential pair.

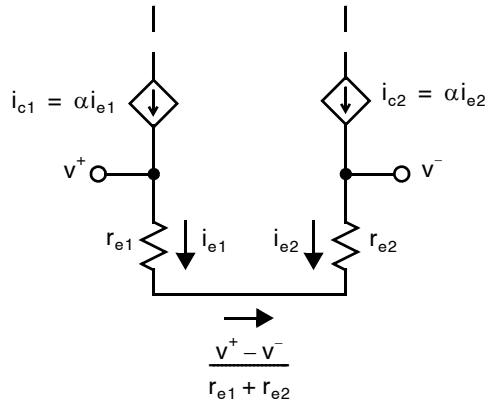


Fig. 8.22 The small-signal model of a bipolar differential pair.

Finally, the input impedance of this differential pair can be found using the impedance-scaling rule just discussed. Specifically, consider the small-signal model shown in Fig. 8.23, where v^- is grounded and we wish to find the impedance r_{id} looking into the base of Q_1. The impedance, r_2, seen looking into the emitter of Q_2 is simply r_{e2}, since the base is grounded. Thus, the total emitter resistance to the ground of Q_1 is $r_{e1} + r_{e2}$, and using the impedance reflection rule, the impedance, r_{id}, seen looking into the base of Q_1 is equal to $\beta + 1$ times the emitter resistance, or, equivalently,

$$r_{id} = (\beta + 1)(r_{e1} + r_{e2}) \tag{8.84}$$

High-Frequency

A symetric bipolar differential amplifier with resistive loads is shown in Fig. 8.24. A small-signal model of this amplifier using the T-model is shown in Fig. 8.25. Here we have ignored both the base resistances and the transistor output resistances in order to simplify the circuit. Neither of these simplifications is a significant limitation of the analysis which follows since the former could be accounted for by taking R_{s1} and R_{s2} larger, and the latter could be accounted for by taking R_{c1} and R_{c2} smaller.

Assuming $R_{c1} = R_{c2}$, $R_{s1} = R_{s2}$, and that the transistors are matched, then the circuit is perfectly symetric. If one also assumes that $v_{in}^- = -v_{in}^+$, then based on the symetry, the node voltage v_e will never change similar to if it was connected to a small-signal ground. Indeed, if the node v_e was connected to a small-signal ground, then circuit will operate identically. This allows one to simplify the analysis to that of the *half-circuit;* with the half-circuit being identical to that of a common-emitter amplifier whose analysis is very similar to that of the common-source MOS amplifier studied previously. Given this equivalence, and assuming the input pole is dominant, the amplifier bandwidth, ω_{-3dB}, is given by

$$\omega_{-3dB} = \frac{1}{(R_{s1} \parallel r_{\pi 1})[C_{be1} + C_{cb1}(1 + g_{m1}R_{c1})]} \tag{8.85}$$

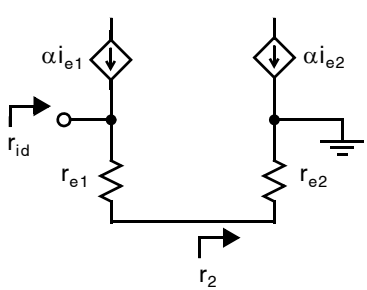


Fig. 8.23 Finding the input impedance of a bipolar differential pair.

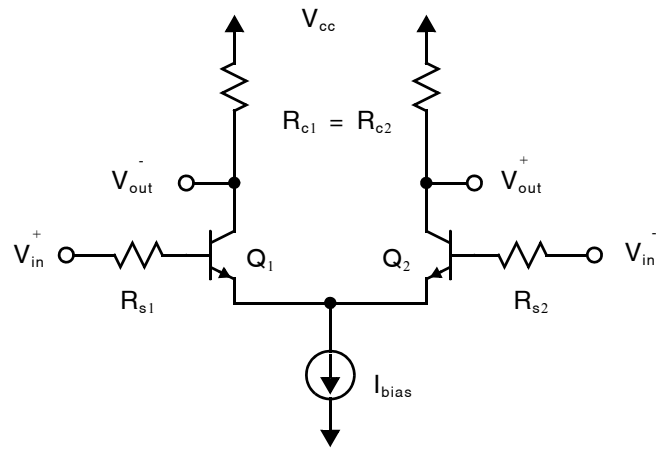


Fig. 8.24 Resistively-loaded symetric BJT differential amplifier.

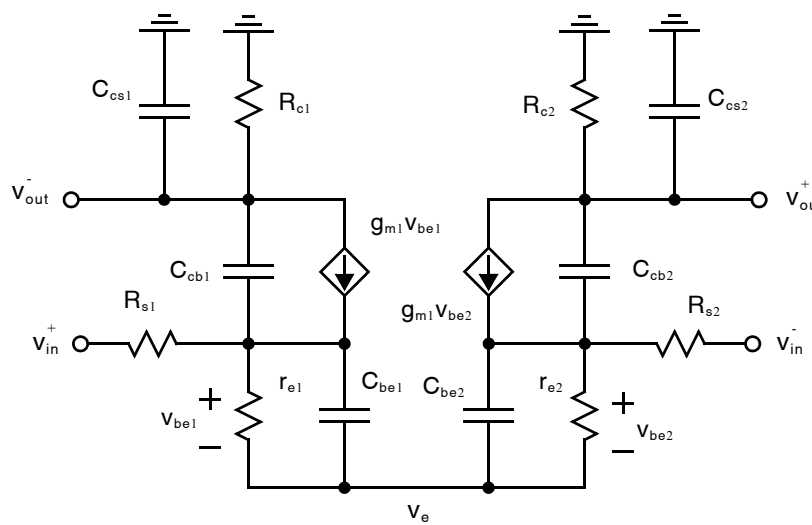


Fig. 8.25 The small-signal model of the resistively-loaded BJT differential amplifier in Fig. 8.24.

Note the important role of the Miller effect which applies a multiplicative gain of $(1 + g_{m1}R_{c1})$ to the collector-base capacitance C_{cb1}. A second pole arises at the output, $\omega_{p2} \cong 1/R_{c1}C_{cs1}$.

8.6 APPENDIX

8.6.1 Bipolar Transistor Exponential Relationship

The various components of the base, collector, and emitter currents were shown in Fig. 8.4, on page 333. Figure 8.26 shows plots of the minority-carrier concentrations in the emitter, base, and collector regions. The current flow of these minority carriers is due to diffusion. By calculating the gradient of the minority-carrier concentrations near the base-emitter junction in a manner similar to that used for diodes, it is possible to derive a relationship between the electron current and the hole current of Fig. 8.4.

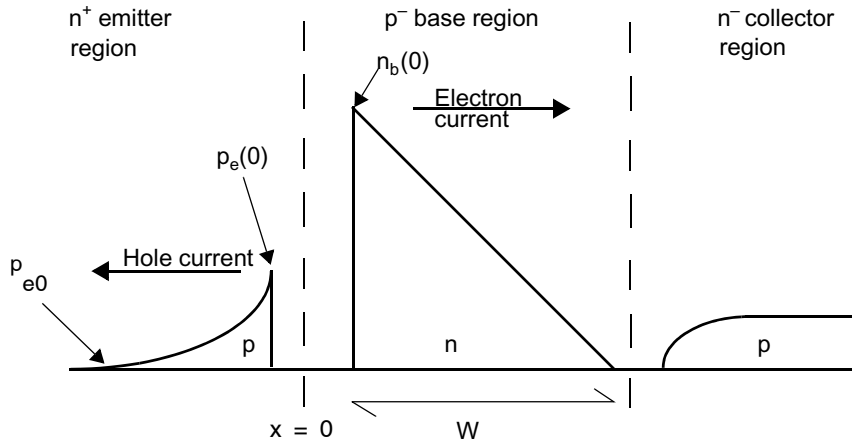


Fig. 8.26 The concentrations of minority carriers in the emitter, base, and collector.

The concentration of holes in the emitter at the edge of the base-emitter depletion region is denoted $p_e(0)$. This concentration decreases exponentially the farther one gets from the junction, in a manner similar to that described for diodes. The concentration far from the junction, p_{e0}, is given by

$$p_{e0} = \frac{n_i^2}{N_D} \tag{8.86}$$

where N_D is the doping density of the n^+ emitter. At a distance x from the edge of the emitter-base depletion region, we have

$$p_e(x) \cong p_e(0)e^{-x/L_p} \tag{8.87}$$

where

$$\begin{aligned} p_e(0) &= p_{e0}e^{V_{BE}/V_T} \\ &= \frac{n_i^2}{N_D}e^{V_{BE}/V_T} \end{aligned} \tag{8.88}$$

and where V_{BE} is the forward-bias voltage of the base-emitter junction.

At the edge of the base-emitter depletion region, the gradient of the hole concentration in the emitter is found, using (8.87), to be

$$\left.\frac{dp_e(x)}{dx}\right|_{x=0} = \frac{p_e(0)}{L_p} \tag{8.89}$$

Using (8.88), we can rewrite this as

$$\left.\frac{dp_e(0)}{dx}\right|_{x=0} = \frac{n_i^2}{L_pN_D}e^{V_{BE}/V_T} \tag{8.90}$$

The hole current is now found using the diffusion equation

$$I_{pe} = A_E q D_p \left.\frac{dp_e(x)}{dx}\right|_{x=0} \tag{8.91}$$

where A_E is the effective area of the emitter. Recall that the minority-hole current in the emitter, I_{pe}, is closely equal to the base current, I_B. After combining (8.90) and (8.91), we obtain

$$I_B = \frac{A_E q D_p n_i^2}{L_p N_D} e^{v_{BE}/V_T} \tag{8.92}$$

The situation on the base side of the base-emitter junction is somewhat different. The concentration of the minority carriers, in this case electrons that diffused from the emitter, is given by a similar equation,

$$n_b(0) = \frac{n_i^2}{N_A} e^{v_{BE}/V_T} \tag{8.93}$$

However, the gradient of this concentration at the edge of the base-emitter depletion region is calculated differently. This difference in gradient concentration is due to the close proximity of the collector-base junction, where the minority carrier (electron) concentration, $n_b(W)$, must be zero. This zero concentration at the collector-base junction occurs because any electrons diffusing to the edge of the collector-base depletion region immediately drift across the junction to the collector, as stated previously. If the base "width" is much shorter than the diffusion length of electrons in the base, L_n, then almost no electrons will recombine with base majority carriers (holes) before they diffuse to the collector-base junction. Given this fact, the decrease in electron or minority concentration from the base-emitter junction to the collector-base junction is a linear relationship decreasing from $n_b(0)$ at the emitter junction to zero at the collector junction in distance W. This assumption ignores any recombination of electrons in the base as they travel to the collector, which is a reasonable assumption for modern transistors that have very narrow bases. Thus, throughout the base region, the gradient of the minority-carrier concentration is closely given by

$$\begin{aligned} \frac{dn_b(x)}{dx} &= -\frac{n_b(0)}{W} \\ &= -\frac{n_i^2}{WN_A} e^{v_{BE}/V_T} \end{aligned} \tag{8.94}$$

Combining (8.94) with the diffusion equation, we obtain

$$\begin{aligned} I_{nb} &= -A_E q D_n \frac{dn_b(0)}{dx} \\ &= \frac{A_E q D_n n_i^2}{WN_A} e^{v_{BE}/V_T} \end{aligned} \tag{8.95}$$

Remembering that I_{nb} is closely equal to the collector current I_C, we have

$$I_C \cong I_{CS} e^{v_{BE}/V_T} \tag{8.96}$$

where

$$I_{CS} = \frac{A_E q D_n n_i^2}{WN_A} \tag{8.97}$$

The ratio of the collector current to the base current, commonly called the *transistor common-emitter current gain* and denoted β, is found using (8.96), (8.97), and (8.92). We have

$$\beta = \frac{I_C}{I_B} = \frac{D_n N_D L_p}{D_p N_A W} \cong 2.5 \frac{N_D L_p}{N_A W} \tag{8.98}$$

which is a constant independent of voltage and current. Noting that $N_D >> N_A$, $L_p > W$, and $D_n \cong 2.5\, D_p$, we have $\beta >> 1$. A typical value might be between 50 and 200. The derivation of β just presented ignores many

second-order effects that make β somewhat current and voltage dependent and are beyond the scope of this book. Interested readers should see [Roulston, 1990] for more details. Regardless of second-order effects, equation (8.98) does reflect the approximate relationships among β, doping levels, and base width. For example, (8.98) explains why heavily doped emitters are important to achieve large current gain.

8.6.2 Base Charge Storage of an Active BJT

Figure 8.26 shows a minority-carrier storage in the base region, Q_b, given by

$$Q_b = A_E q \frac{n_b(0) W}{2} \tag{8.99}$$

Using (8.93) for nb(0), we have

$$Q_b = \frac{A_E q n_i^2 W}{2N_A} e^{V_{BE}/V_T} \tag{8.100}$$

This equation can be rewritten using (8.96) and (8.97) to obtain

$$Q_b = \frac{W^2}{2D_n} I_C = \tau_b I_C \tag{8.101}$$

where τ_b, called the *base-transit time constant,* is given approximately by

$$\tau_b = \frac{W^2}{2D_n} \tag{8.102}$$

ignoring second-order effects. Normally, the base-transit time constant is specified for a given technology and takes into account other charge-storage effects not considered here, and is therefore often denoted τ_T. However, since the base storage of electrons dominates the other effects, we have $\tau_T \cong \tau_b$.

If the current in a BJT changes, the base charge storage must also change. This change can be modelled by a diffusion capacitance, C_d, between the base and the emitter terminals. Using (8.101), we have

$$C_d \cong \frac{dQ_b}{dV_{BE}} = \frac{d(\tau_b I_C)}{dV_{BE}} \tag{8.103}$$

Using (8.103) and $I_C = I_{CS} e^{V_{BE}/V_T}$ results in

$$C_d = \tau_b \frac{I_C}{V_T} \tag{8.104}$$

This equation is similar to that for a diode.

8.7 SUMMARY OF KEY POINTS

- Bipolar devices offer advantages over CMOS devices when high-speed operation must be combined with high gain and/or high breakdown voltages. However, unlike MOSFETs, they draw a nonzero dc base current. [p. 331]
- In active mode, with base-emitter junction forward-biased and base-collector junction reverse-biased, collector current increases exponentially with base-emitter voltage, and the base current is a factor smaller. The variation in collector current with collector voltage is first-order modeled by the Early voltage. [p. 334]
- As the base-collector junction approaches forward-bias, at collector-emitter voltages of about 0.2 to 0.3 V, a finite base-collector current flows, and the transistor is saturated. [p. 335]

- The time required to remove the charge stored in a saturated transistor is much larger than the time required to turn off a transistor in the active region. Hence, in high-speed microcircuit designs, one never allows bipolar transistors to saturate. [p. 337]
- BJT small-signal transconductance is proportional to collector current. [p. 338]
- The high-frequency operation of active BJTs is limited by capacitances modeling charge storage in the depletion regions and the storage of diffusing charge-carriers in the base. [p. 341]
- Active BJTs offer an intrinsic gain that is generally higher than that of MOS transistors and independent of operating point. [p. 342]
- BJT unity-gain frequency increases in proportion to bias current up to a limit imposed by the transit time of carriers through the base. Hence, the base thickness is a critical parameter limiting the maximum operating speed of a BJT. [p. 343]
- High-performance modern bipolar transistors are manufactured by epitaxially growing a very thin base region. The use of a different material for the base, such as silicon-germanium, results in a heterojunction bipolar transistor (HBT) whose bandgap structure can be tailored to provide high current gain, high unity-gain frequency, and/or high Early voltage. [p. 348]
- Bipolar current mirrors are similar to CMOS mirrors, except that finite base current is a source of mismatch between the input and output currents. [p. 349]
- At low frequencies, resistances in series with the emitter appear (β+1) times larger when seen looking into the base. [p. 352]
- Resistances in series with the base are divided by (β+1) when they are seen looking into the emitter. [p. 353]

8.8 REFERENCES

Harame, D. L., Comfort, J. H., Cressler, J .D., Crabbe, E. F., Sun, J. Y.-C., Meyerson, B. S., Tice, T., "Si/SiGe epitaxial-base transistors. I. Materials, physics, and circuits," *IEEE Transactions on Electron Devices*, vol. 42, no. 3, pp.455–468, March 1995.

Harame, D. L., Comfort, J. H., Cressler, J. D., Crabbe, E. F., Sun, J. Y.-C., Meyerson, B. S., Tice, T., "Si/SiGe epitaxial-base transistors. II. Process integration and analog applications," *IEEE Transactions on Electron Devices*, vol. 42, no. 3, pp.469–482, March 1995.

D. Hodges and H. Jackson. *Analysis and Design of Digital Integrated Circuits,* 2nd ed. McGraw-Hill, New York, 1988.

Kirk, C. T., Jr., "A theory of transistor cutoff frequency (f_T) falloff at high current densities," *IRE Transactions on Electron Devices*, vol. 9, no. 2, pp.164–174, March 1962.

D. Roulston. *Semiconductor Devices.* McGraw-Hill, New York, 1990.

S. M. Sze. *Physics of Semiconductor Devices.* Wiley Interscience, New York, 1981.

Tuinhout, H. P., "Improving BiCMOS technologies using BJT parametric mismatch characterisation," *Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 163–170, Sept. 2003.

8.9 PROBLEMS

Unless otherwise stated, assume npn bipolar transistors with parameters as listed in Table 8.1.

8.9.1 Section : 8.1 Bipolar-Junction Transistors

8.1 For an npn transistor having $I_C = 0.1\ mA$, calculate g_m, r_π, r_e, r_0, and $g_m r_0$.

8.2 Initially, the circuit shown in Fig. P8.2 has a 0-V input. At time 0 its input changes to 5 V. Estimate the time it takes its output voltage to saturate, using the concepts of average capacitance and first-order transient solutions for each node. The time constants of the individual nodes can be added to arrive at an overall time constant for the approximate first-order transient response of the circuit. Next, assume that the input changes from 5 V to 0 V at time 0. How long does it take the output voltage to change to 3.5 V?

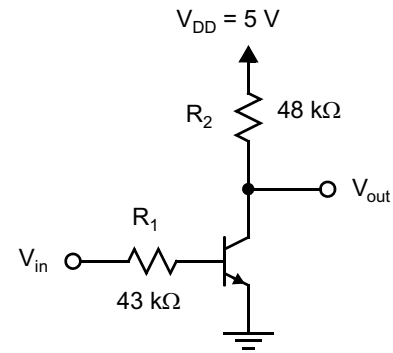


Fig. P8.2

8.3 Compare your answers to Problem 8.2 to those obtained using SPICE.

8.4 Verify that for $I_{BR} >> I_B$ and $I_{BR} >> I_C / \beta$, Eq. (8.14) simplifies to Eq. (8.13).

8.9.2 Section : 8.5 Bipolar Current Mirrors and Gain Stages

8.5 Derive the current gain, I_{out}/I_{in}, for the bipolar current mirror shown in Fig. 8.13, and show that, for $\beta >> 1$, this gain is $1 - 2/\beta$. (Neglect finite output impedances.)

8.6 Derive the current gain, I_{out}/I_{in}, for the bipolar current mirror shown in Fig. 8.14, and show that, for $\beta >> 1$, this gain is $1 - 2/\beta^2$. (Neglect finite output impedances.)

8.7 For the bipolar cascode current mirror shown in Fig. 8.16(*a*), derive the current gain, I_{out}/I_{in}, and show that, for large β, this gain is given by

$$\frac{I_{out}}{I_{in}} \approx 1 - \frac{4}{\beta}$$

8.8 For the bipolar Wilson current mirror shown in Fig. 8.16(*b*), show that the current gain, I_{out}/I_{in}, is given by

$$\frac{I_{out}}{I_{in}} = 1 - \frac{2}{\beta^2 + 2\beta + 2}$$

8.9 Derive the output impedance of the simple bipolar current mirror shown in Fig. P8.9. Take the finite output impedance, r_o, into account for both transistors. What is the required value for R_{bias}, and what is the output impedance if we want the transistors to be biased at 0.2 mA ?

8.10 Derive the –3-dB frequency of a bipolar common-emitter amplifier, such as the one illustrated in Fig. P8.2, in terms of R_1, R_2, the operating point collector current, and transistor parameters.

8.11 Derive the low-frequency gain and output impedance of a bipolar emitter follower that is biased with a current mirror. What are these values for a 0.5-mA bias current?

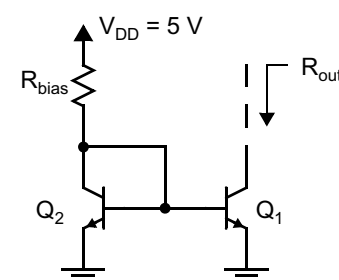


Fig. P8.9

8.12 Derive the low-frequency input impedance and the gain of the bipolar common-base amplifier illustrated in Fig. P8.12.

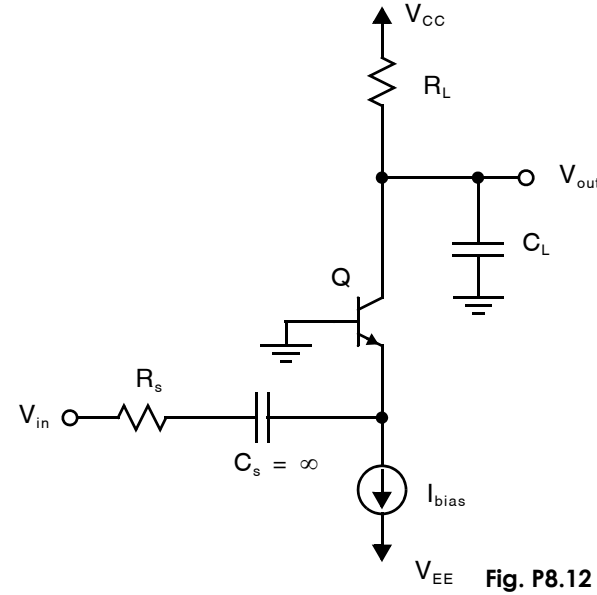


Fig. P8.12

8.13 For the circuit shown in Fig. P8.13, show that the output impedance, r_{out}, is approximately given by

$$r_{out} \cong r_{o2}(1 + g_{m2}(R_e \parallel r_\pi))$$

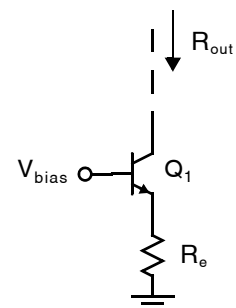


Fig. P8.13

8.14 Taking into account the finite current gain of bipolar transistors, but ignoring the finite output impedance, find the large-signal current gains of cascode and Wilson current mirrors.

8.15 Assuming $\beta >> 1$ and $g_m r_o >> \beta$, show that the output impedance of a bipolar Wilson current mirror is approximately given by

$$r_{out} \cong \frac{\beta r_o}{2}$$

8.16 Repeat Problem 8.15 for a bipolar cascode current mirror.

8.17 Find an expression for the frequency response of the single-ended differential amplifier in Fig. P8.17 in terms of the component values and transistor small-signal parameters. Show that its bandwidth is approximately $1/(R_s \parallel 2r_\pi)(C_{be}/2 + C_{cb1})$.

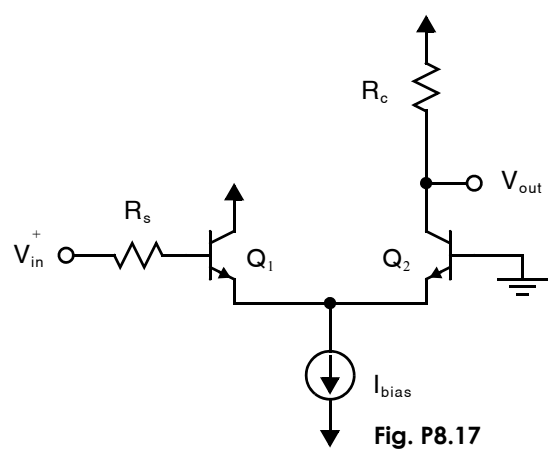


Fig. P8.17

CHAPTER

9 Noise and Linearity Analysis and Modelling

To develop good analog circuit design techniques, a basic understanding of noise sources and analysis is required. Another motivation to study noise analysis is to learn basic concepts of random signals for a proper understanding of oversampling converters. The purpose of this chapter is to present some fundamentals of noise analysis followed by an introduction to electronic noise sources and circuit analysis, and finally linearity.

It should be mentioned here that this chapter deals with *inherent noise* as opposed to *interference noise*. Interference noise is a result of unwanted interaction between the circuit and the outside world, or between different parts of the circuit itself. This type of noise may or may not appear as random signals. Examples are power supply noise on ground wires (such as a 60-Hz hum) or electromagnetic interference between wires. Interference noise can be significantly reduced by careful circuit wiring or layout. In contrast, inherent noise refers to random noise signals that can be reduced but never eliminated since this noise is due to fundamental properties of the circuits. Some examples of inherent noise are thermal, shot, and flicker noise. Inherent noise is only moderately affected by circuit wiring or layout, such as using multiple base contacts to change the resistance value of the base of a transistor. However, inherent noise can be significantly reduced through proper circuit design, such as changing the circuit structure or increasing the power consumption.

The outline of this chapter is as follows: First, a time-domain analysis of noise signals (or other random signals) is presented. Here, basic concepts such as rms value, signal-to-noise ratio, and noise summation are presented. Next, a frequency-domain analysis of noise is presented. As with deterministic signals, analysis in the frequency domain results in more powerful analysis tools than does analysis that remains strictly in the time domain. Noise models for circuit elements are then presented, followed by noise analyses of some real circuits to give the reader some experience in such analysis. Finally, whereas noise imposes a lower limit on the range of signal amplitudes that can be meaningfully processed by a circuit, linearity often imposes the upper limit. The difference between them is a circuit's *dynamic range*. The chapter ends with introductions to linearity and dynamic range.

9.1 TIME-DOMAIN ANALYSIS

Since inherent noise signals are random in nature, we define here some basic tools to effectively deal with random signals. Specifically, in this section, we define the following terms in the time domain: rms value, SNR, dBm, and noise summation.

> ***Key Point:*** *This chapter focuses on noise inherent to the physical operation of electronic components, not noise from outside interferences. Such noise sources are generally zero-mean.*

An example of a random noise signal in the time domain is shown in Fig. 9.1. Although this signal is a voltage it could just as easily be current noise or some other quantity. The signal appears to have an average value of zero. In fact, *throughout this chapter we will assume all noise signals have a mean value of zero*, which simplifies many of the definitions and is also valid in most physical systems.

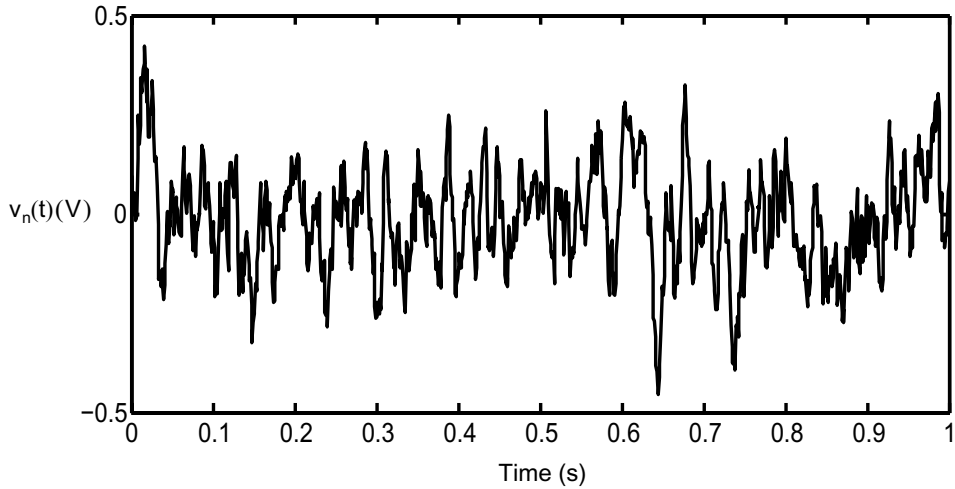


Fig. 9.1 Example of a voltage noise signal in the time domain.

9.1.1 Root Mean Square (rms) Value

Consider a noise voltage, $v_n(t)$, such as that shown in Fig. 9.1, or a noise current, $i_n(t)$. The *rms*, or *root mean square*, voltage value is defined[1] as

$$V_{n(rms)} \equiv \left[\frac{1}{T}\int_0^T v_n^2(t)\ dt\right]^{1/2} \tag{9.1}$$

where T is a suitable averaging time interval. Typically, a longer T gives a more accurate rms measurement. Similarly, the rms current value is defined as

$$I_{n(rms)} \equiv \left[\frac{1}{T}\int_0^T i_n^2(t)\ dt\right]^{1/2} \tag{9.2}$$

The benefit in knowing the rms value of a signal is that it indicates the *normalized noise power* of the signal. Specifically, if the random signal $v_n(t)$ is applied to a 1-Ω resistor, the average power dissipated, P_{diss}, in watts, equals the normalized noise power and is given by

$$P_{diss} = \frac{V_{n(rms)}^2}{1\ \Omega} = V_{n(rms)}^2 \tag{9.3}$$

Key Point: *Although the instantaneous value of a noise voltage or current cannot be determined by analysis, the rms value obtained by time-averaging its square provides a normalized measure of the noise signal's power.*

This relationship implies that the power dissipated by a resistor is the same whether a random signal or a dc level of k volts (rms) is applied across it. For example, a noise signal with an rms value of 1 mV (rms) dissipates the same power across a resistor as a dc voltage of 1 mV.

Similarly, for a noise current, $i_n(t)$, applied to a 1-Ω resistor,

$$P_{diss} = I_{n(rms)}^2 \times 1\ \Omega = I_{n(rms)}^2 \tag{9.4}$$

As a result, the square of the rms values, $V_{n(rms)}^2$ and $I_{n(rms)}^2$, are sometimes referred to as the *normalized noise powers* of these two signals.

1. For those more rigorously inclined, we assume throughout this chapter that random signals are ergodic, implying their ensemble averages can be approximated by their time averages.

9.1.2 SNR

The *signal-to-noise ratio (SNR)* value (in dB) of a signal node in a system is defined as

$$\text{SNR} \equiv 10 \log \left[\frac{\text{signal power}}{\text{noise power}} \right] \text{dB} \tag{9.5}$$

Thus, assuming a node in a circuit consists of a signal, $v_x(t)$, that has a normalized signal power of $V_{x(\text{rms})}^2$ and a normalized noise power of $V_{n(\text{rms})}^2$, the SNR is given by

$$\text{SNR} = 10 \log \left[\frac{V_{x(\text{rms})}^2}{V_{n(\text{rms})}^2} \right] = 20 \log \left[\frac{V_{x(\text{rms})}}{V_{n(\text{rms})}} \right] \text{dB} \tag{9.6}$$

Key Point: *SNR is the ratio of signal power to noise power, both normalized to the same impedance and expressed in dB.*

Clearly, when the mean-squared values of the noise and signal are the same, then $\text{SNR} = 0$ dB.

9.1.3 Units of dBm

Although dB units relate the relative ratio of two power levels, it is often useful to know a signal's power in dB on an absolute scale. One common measure is that of dBm, where all power levels are referenced by 1 mW. In other words, a 1-mW signal corresponds to 0 dBm, whereas a 1-μW signal corresponds to -30 dBm. When voltage levels are measured, it is also common to reference the voltage level to the equivalent power dissipated if the voltage is applied across either a 50-Ω or a 75-Ω resistor.

Key Point: *Units of dBm express an average power level on a logarithmic scale normalized to 1 mW, so that 1 mW is 0 dBm and 0.1 mW is -10 dBm. Rms voltages or currents are also sometimes expressed in dBm by assuming that they are applied to some agreed-upon reference load, usually either 50 or 75 Ohms.*

EXAMPLE 9.1

Find the rms voltage of a 0-dBm signal referenced to a 50-Ω resistor. What is the level in dBm of a 2-volt rms signal?

Solution

A 0-dBm signal referenced to a 50-Ω resistor implies that the rms voltage level equals

$$V_{(\text{rms})} = \sqrt{(50\ \Omega) \times 1\ \text{mW}} = 0.2236\ \text{V} \tag{9.7}$$

Thus, a 2-volt rms signal corresponds to

$$20 \log \left(\frac{2.0}{0.2236} \right) = 19\ \text{dBm} \tag{9.8}$$

and would dissipate 80 mW across a 50-Ω resistor.

Note that the measured voltage may never be physically applied across any 50-Ω resistor. The measured voltage is referenced only to power levels that would occur if the voltage were applied to such a load. Similar results are obtained if the power is referenced to a 75-Ω resistor.

9.1.4 Noise Summation

Consider the case of two noise sources added together, as shown in Fig. 9.2. If the rms values of each individual noise source are known, what can be said about the rms value of the combined signal? We answer this question as follows. Define $v_{no}(t)$ as

$$v_{no}(t) = v_{n1}(t) + v_{n2}(t) \tag{9.9}$$

where $v_{n1}(t)$ and $v_{n2}(t)$ are two noise sources with known rms values $V_{n1(rms)}$ and $V_{n2(rms)}$, respectively. Then we can write

$$V_{no(rms)}^2 = \frac{1}{T}\int_0^T [v_{n1}(t) + v_{n2}(t)]^2 \, dt \tag{9.10}$$

which, when expanded, gives,

$$V_{no(rms)}^2 = V_{n1(rms)}^2 + V_{n2(rms)}^2 + \frac{2}{T}\int_0^T v_{n1}(t)v_{n2}(t) \, dt \tag{9.11}$$

Note that the first two terms in the right-hand side of (9.11) are the individual mean-squared values of the noise sources. The last term shows the correlation between the two signal sources, $v_{n1}(t)$ and $v_{n2}(t)$. An alternate way to write (9.11) that better indicates the effects of signal correlation, is to define a correlation coefficient, C, as

$$C \equiv \frac{\frac{1}{T}\int_0^T v_{n1}(t)v_{n2}(t) \, dt}{V_{n1(rms)}V_{n2(rms)}} \tag{9.12}$$

With this definition, (9.11) can also be written as

$$V_{no(rms)}^2 = V_{n1(rms)}^2 + V_{n2(rms)}^2 + 2CV_{n1(rms)}V_{n2(rms)} \tag{9.13}$$

It can be shown that the correlation coefficient always satisfies the condition $-1 \le C \le 1$. Also, a value of $C = \pm 1$ implies the two signals are fully correlated, whereas $C = 0$ indicates the signals are uncorrelated. Values in between imply the signals are partially correlated. Fortunately, we have little reason to analyze partially correlated signals since different inherent noise sources are typically uncorrelated.

In the case of two uncorrelated signals, the mean-squared value of their sum is given by

$$V_{no(rms)}^2 = V_{n1(rms)}^2 + V_{n2(rms)}^2 \tag{9.14}$$

Key Point: *The power of independent (uncorrelated) noise sources superimpose linearly, which means their voltages and currents superimpose in a root-sum-of-squares fashion. As a result, to lower overall noise, one must focus on the largest noise contributors.*

This relationship indicates that two rms values add as though they were the magnitudes of vectors at right angles to each other (i.e., orthogonal) when signals are uncorrelated.

It is of interest to contrast this uncorrelated case with that for fully correlated signals. An example of two fully correlated (though deterministic) sources are two sinusoidal signals that have the same frequency and a phase of 0 or 180 degrees with each other. In this fully correlated case, the mean-squared value of their sum is given by

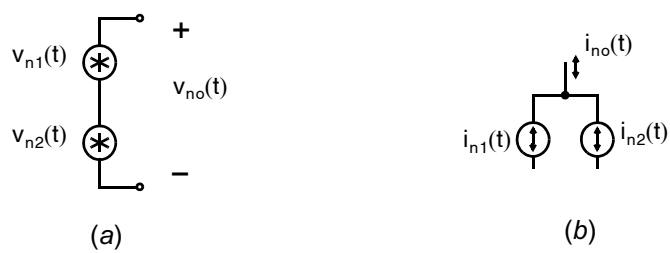


Fig. 9.2 Combining two noise sources, (*a*) voltage, and (*b*) current.

$$V_{no(rms)}^2 = [V_{n1(rms)} \pm V_{n2(rms)}]^2 \tag{9.15}$$

where the sign is determined by whether the signals are in or out of phase with each other. Note that, in this case where the signals are fully correlated, the rms values add linearly (similar to the magnitudes of aligned vectors).

EXAMPLE 9.2

Given two uncorrelated noise sources that have $V_{n1(rms)} = 10\ \mu V$ and $V_{n2(rms)} = 5\ \mu V$, find their total output rms value when combined. If we are required to maintain the total rms value at $10\ \mu V$, how much should $V_{n1(rms)}$ be reduced while $V_{n2(rms)}$ remains unchanged?

Solution

Using (9.14) results in

$$V_{no(rms)}^2 = (10^2 + 5^2) = 125(\mu V)^2 \tag{9.16}$$

which results in $V_{no(rms)} = 11.2\ \mu V$.

To maintain $V_{no(rms)} = 10\ \mu V$ and $V_{n2(rms)} = 5\ \mu V$, we have

$$10^2 = V_{n1(rms)}^2 + 5^2 \tag{9.17}$$

which results in $V_{n1(rms)} = 8.7\ \mu V$. Therefore, reducing $V_{n1(rms)}$ by 13 percent is equivalent to eliminating $V_{n2(rms)}$ altogether!

The above example has an important moral. *To reduce overall noise, concentrate on large noise signals*.

9.2 FREQUENCY-DOMAIN ANALYSIS

As with deterministic signals, frequency-domain techniques are useful for dealing with random signals such as noise. This section presents frequency-domain techniques for dealing with noise signals and other random signals. It should be noted that units of hertz (Hz) (rather than radians/second) are used throughout this chapter since, historically, such units have been commonly used in the measurement of continuous-time spectral densities.

9.2.1 Noise Spectral Density

Although periodic signals (such as a sinusoid) have power at distinct frequency locations, random signals have their power spread out over the frequency spectrum. For example, if the time-domain signal shown in Fig. 9.1 is applied to a spectrum analyzer, the resulting spectrum might look like that shown in Fig. 9.3(*a*). Note here that although the horizontal scale is the usual frequency axis, the vertical scale is in units of microvolts-squared/hertz. In other words, the vertical axis is a measure of the normalized noise power (mean-squared value) over a 1-Hz bandwidth at each frequency point. For example, the measurement at 100 Hz in Fig. 9.3(*a*) indicates that the normalized power between 99.5 Hz and 100.5 Hz is $10\ (\mu V)^2$.

Thus, we define the noise spectral density, $V_n^2(f)$, or in the case of current, $I_n^2(f)$, as the average normalized noise power over a 1-Hz bandwidth. The units of $V_n^2(f)$ are volts-squared/hertz, whereas those of $I_n^2(f)$ are amps-squared/hertz. Also, $V_n^2(f)$ is a positive real-valued function.

Key Point: *The frequency-domain representation of a noise signal is its spectral density: the average normalized power that may be measured within a 1-Hz bandwidth.*

It should be emphasized here that the *mean-squared value of a random noise signal at a single precise frequency is zero*. In other words, the mean-squared value of the signal shown in Fig. 9.3(*a*) measured at 100 Hz is directly proportional to the bandwidth of the bandpass filter used for the measurement. In a laboratory spectrum analyzer, the bandwidth of the bandpass filter is determined by the resolution-bandwidth control. Thus, as the resolution bandwidth goes to zero, the mean-squared value also becomes zero.[2] Conversely, as the resolution bandwidth increases, so does the measured mean-squared value. In either case, the measured mean-squared value should be normalized to the value that would be obtained for a bandwidth of 1 Hz when the noise spectral density is stated in units of $V^2/(Hz)$. For example, if the signal corresponding to Fig. 9.3(*a*) were measured at around 0.1 Hz using a resolution bandwidth of 1 mHz, the mean-squared value measured would be 1 $(\mu V)^2$, which, when scaled to a 1-Hz bandwidth, equals 1,000 $(\mu V^2)/Hz$.

An intuitive explanation of how a random noise signal is measured using a spectrum analyzer is as follows. A random noise signal has a frequency that is continually changing through a broad continuum of frequencies. A spectrum analyzer is sensitive to only a narrow frequency range that is the passband of its filter, and it measures the mean-squared value in that range. The filter of the spectrum analyzer reacts with a time constant approximately given by

$$\tau \approx \frac{1}{\pi W} \tag{9.18}$$

where W is the bandwidth of the filter. For some of the time, the random signal has the same frequency as the spectrum analyzer's filter. Thus, the spectrum analyzer effectively measures what percentage of time the random signal is in the frequency range of its filter. The narrower the bandwidth of the filter, the less percentage of time the signal is within its frequency range, and therefore the smaller is the spectrum analyzer's output. However, if the signal is not a random signal wandering in and out of the frequency range of the spectrum analyzer's filter, but is a deterministic signal at the center frequency of the filter, then the spectrum analyzer's reading is independent of the filter's bandwidth.

It is often convenient to plot the square root of the noise spectral density when we deal with filtered noise. Taking a square root results in $V_n(f)$, as shown in Fig. 9.3(*b*). We will refer to $V_n(f)$ as the *root spectral density which is expressed in units of volts/root-hertz* (i.e., $V/\sqrt{Hz}$). In the case of current noise, the resulting units are *amps/root-hertz*. Note that the horizontal axis remains unchanged although there is a root-hertz factor in the vertical axis.

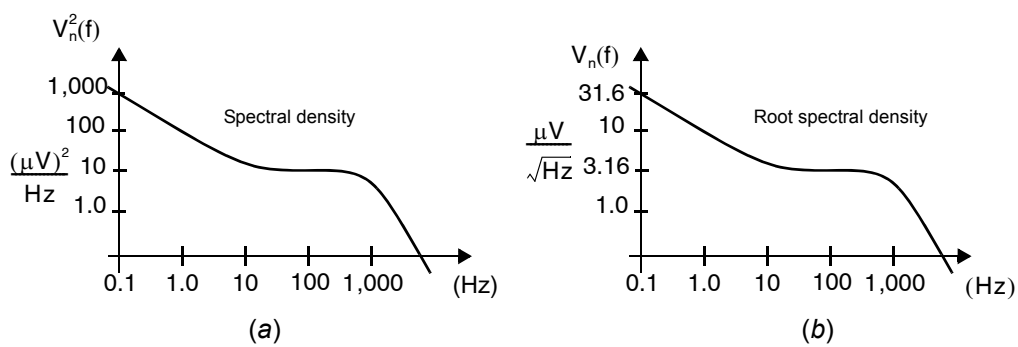


Fig. 9.3 Example of voltage spectral density (frequency domain), for (*a*) spectral density, and (*b*) root spectral density.

2. Such a result would not occur when a 100-Hz sinusoidal waveform is measured.

Since the spectral density measures the mean-squared value over a 1-Hz bandwidth, one can obtain the total mean-squared value by integrating the spectral density over the entire frequency spectrum. Thus, the rms value of a noise signal can also be obtained in the frequency domain using the following relationship:

Key Point: *The mean square value of a noise signal is the integral of its spectral density over all frequencies.*

$$V_{n(rms)}^2 = \int_0^\infty V_n^2(f)\,df \tag{9.19}$$

and similarly for current noise,

$$I_{n(rms)}^2 = \int_0^\infty I_n^2(f)\,df \tag{9.20}$$

More generally, the rms noise within a specified frequency range is obtained by integrating the noise spectral density over that frequency range. For example, the rms voltage noise over the frequency range $f_1 < f < f_2$ is

$$\int_{f_1}^{f_2} V_n^2(f)\,df \tag{9.21}$$

Finally, $V_n^2(f)$ is rigorously defined as the Fourier transform of the autocorrelation function of the time-domain signal, $v_n(t)$. This relationship is known as the *Wiener–Khinchin theorem*. The relationship just shown defines a one-sided spectral density function since the noise is integrated only over positive frequencies as opposed to both negative and positive frequencies, again primarily for historical reasons. A two-sided definition results in the spectral density being divided by two since, for real-valued signals, the spectral density is the same for positive and negative frequencies, and the total mean-squared value obtained by integrating over both positive and negative frequencies remains unchanged.

EXAMPLE 9.3

What mean-squared value would be measured on the signal shown in Fig. 9.3 at 100 Hz when a resolution bandwidth of 30 Hz is used on a spectrum analyzer? Answer the same question for a 0.1-Hz resolution bandwidth.

Solution

Since the portion of spectral density function is flat at about 100 Hz, the measured value should simply be proportional to the bandwidth. Since the noise spectral density is 10 $(\mu V)^2/Hz$ at 100 Hz, the output of a 30-Hz filter is $30\ Hz \times 10\ (\mu V)^2/Hz = 300\ (\mu V)^2$ (or an rms value of $\sqrt{300}\ \mu V$).

For a 0.1-Hz bandwidth, the measured value would be 10 times smaller than for a 1-Hz bandwidth, resulting in a value of 1 $(\mu V)^2$ (or an rms value of 1 μV).

9.2.2 White Noise

One common type of noise is *white noise*. A noise signal is said to be white if its spectral density is constant over a given frequency. In other words, a white noise signal would have a flat spectral density, as shown in Fig. 9.4, where $V_n(f)$ is given by

Key Point: *White noise has a spectral density that is constant over all frequencies.*

$$V_n(f) = V_{nw} \tag{9.22}$$

and V_{nw} is a constant value.

Substituting (9.22) into (9.19) suggests that white noise has infinite power. However, this is not a problem in practice since a finite capacitance is always present to bandlimit the noise and make the integral in (9.19) converge.

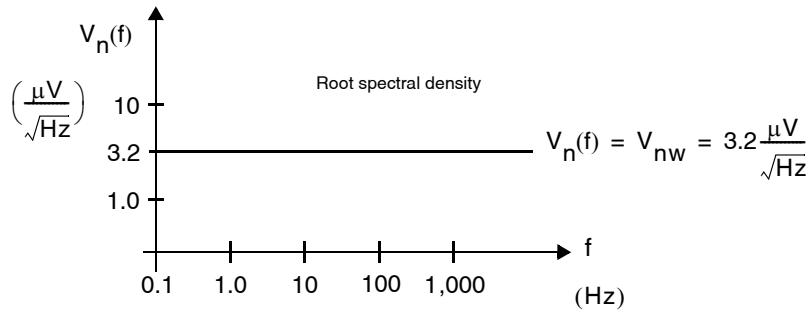


Fig. 9.4 An example of a white noise signal.

9.2.3 1/f, or Flicker, Noise

Another common noise shape is that of *1/f, or flicker,* noise.[3] The spectral density, $V_n^2(f)$, of 1/f noise is approximated by

$$V_n^2(f) = \frac{k_v^2}{f} \tag{9.23}$$

where k_v is a constant. Thus, the *spectral density is inversely proportional to frequency,* and hence the term "1/f noise." In terms of root spectral density, 1/f noise is given by

$$V_n(f) = \frac{k_v}{\sqrt{f}} \tag{9.24}$$

Key Point: *Flicker noise, or 1/f-noise, has a spectral density proportional to 1/f. Hence, it falls off at –10 dB/decade.*

Note that it is inversely proportional to $\sqrt{f}$ (rather than f). An example of a signal having both 1/f and white noise is shown in Fig. 9.5. Note that the 1/f noise falls off at a rate of –10 dB/decade since it is inversely proportional to $\sqrt{f}$. The intersection of the 1/f and white noise curves is often referred to as the *1/f noise corner.* (It occurs at 10 Hz in Fig. 9.5.)

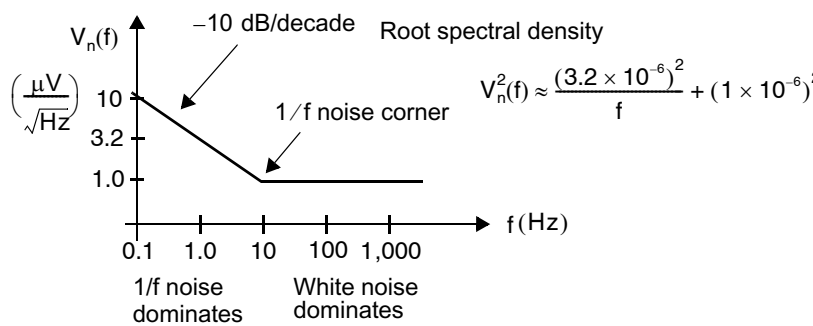


Fig. 9.5 A noise signal that has both 1/f and white noise.

3. 1/f, or flicker, noise is also referred to as *pink noise* since it has a large low-frequency content, like pink light.

Substituting (9.23) into (9.21) gives the power of a 1/f noise source over a finite frequency range $f_1 < f < f_2$.

$$\int_{f_1}^{f_2} \frac{k_v^2}{f}\,df = k_v^2 \ln\left(\frac{f_2}{f_1}\right) \tag{9.25}$$

Hence, the noise power in every decade range of frequencies is equal to $k_v^2 \ln(10) \cong 2.3 k_v^2$. Integration of 1/f noise all the way down to dc yields infinite power, but in practical cases, finite and reasonable values are obtained even when the lower limit of integration is very low.

EXAMPLE 9.4

Find the rms noise voltage of a 1/f noise source with $k_v^2 = (50\ \text{nV}^2)$ integrating over the frequency range 1 Hz to 100 MHz. What is its rms noise voltage if instead the lower limit of integration is decreased to 10 nHz?

Solution

The range 1 Hz to 100 MHz encompasses 8 decades. Hence, the rms noise voltage over that range is $\sqrt{8 \cdot 2.3 \cdot k_v^2} = 0.215\ \mu\text{V}$. Decreasing the lower frequency limit to 10 nHz increases the rms noise voltage to $\sqrt{16 \cdot 2.3 \cdot k_v^2} = 0.303\ \mu\text{V}$. Circuit noise at frequencies around 10 nHz will only be observed after a circuit has been in continuous operation for more than 1 year. In practice, over such a long time, temperature and aging effects will cause shifts in device performance that are virtually indistinguishable from the 1/f noise.

9.2.4 Filtered Noise

Consider the case of a noise signal, $V_{ni}(f)$, being filtered by the transfer function $A(s)$, as shown in Fig. 9.6. Here, $A(s)$ represents a linear transfer function as a result of some circuit amplification, filtering, or both. The following relationship between the input and output signals can be derived using the definition of the spectral density.

$$V_{no}^2(f) = |A(j2\pi f)|^2 V_{ni}^2(f) \tag{9.26}$$

The term $2\pi f$ arises here since, for physical frequencies, we replace s with $j\omega = j2\pi f$. Note that the output spectral density is a function only of the magnitude of the transfer function, and not its phase. As a result of (9.26), the total output mean-squared value is given by

$$V_{no(rms)}^2 = \int_0^\infty |A(j2\pi f)|^2 V_{ni}^2(f)\,df \tag{9.27}$$

If we wish to work with root spectral densities, we can take the square root of both sides of (9.26) resulting in

$$V_{no}(f) = |A(j2\pi f)| V_{ni}(f) \tag{9.28}$$

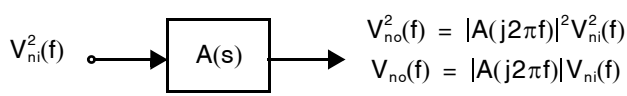


Fig. 9.6 Applying a transfer function (i.e. filter) to a noise signal.

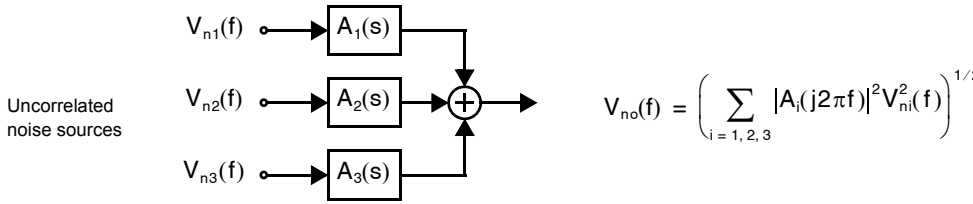


Fig. 9.7 Filtered uncorrelated noise sources contributing to total output noise.

Key Point: *Passing a noise signal through a circuit shapes its root spectral density by the circuit's magnitude response.*

The relationship in (9.28) should make intuitive sense since it indicates that the transfer function simply shapes the root spectral density of the input noise signal. It is important to note here that the root spectral density is simply shaped by $|A(j2\pi f)|$, whereas the spectral density is shaped by $|A(j2\pi f)|^2$, as seen by (9.26). Hence, we see the benefit in dealing with the root spectral density rather than the spectral density. Specifically, straightforward transfer function analysis is applied when using root spectral densities, whereas squared terms are required to deal with spectral densities.

Key Point: *Uncorrelated noise signals remain uncorrelated, even when filtered by a circuit's magnitude response.*

It is also of interest to consider the case of multiple uncorrelated noise sources that are each filtered and summed together. For example, consider the system shown in Fig. 9.7 in which three filtered, uncorrelated noise sources combine to form the total output noise, $V_{no}(f)$. In this case, one can show that if the input random signals are uncorrelated, the filter outputs are also uncorrelated. As a result, the output spectral density is given by

$$V_{no}^2(f) = \sum_{i=1,2,3} |A_i(j2\pi f)|^2 V_{ni}^2(f) \tag{9.29}$$

EXAMPLE 9.5

Consider a noise signal, $V_{ni}(f)$, that has a white root spectral density of 20 $nV/\sqrt{Hz}$, as shown in Fig. 9.8(*a*). Find the total noise rms value between dc and 100 kHz. What is the total noise rms value if it is filtered by the RC filter shown in Fig. 9.8(*b*), where it is assumed the RC filter is noise free?

Solution

For the noise mean-square value from dc to 100 kHz of $V_{ni}(f)$, we have

$$V_{ni(rms)}^2 = \int_0^{10^5} 20^2 \, df = 4 \times 10^7 (nV)^2 \tag{9.30}$$

resulting in an rms value of $V_{ni(rms)} = 6.3$ μV rms . Note that, for this simple case, one could also obtain the rms value by multiplying 20 $nV/\sqrt{Hz}$ by the square root of the frequency span, or $\sqrt{100 \text{ kHz}}$, resulting in 6.3 μV rms .

For the filtered signal, $V_{no}(f)$, we find that the RC filter has the frequency response shown in Fig. 9.8(*c*). Therefore, we can multiply the root spectral density of $V_{ni}(f)$ with the frequency response, $|A(j2\pi f)|$, to obtain the root spectral density of $V_{no}(f)$, as shown in Fig. 9.8(*d*). Mathematically, the output root spectral density is given by

$$V_{no}(f) = \frac{20 \times 10^{-9}}{\sqrt{1 + \left(\frac{f}{f_0} \right)^2}} \tag{9.31}$$

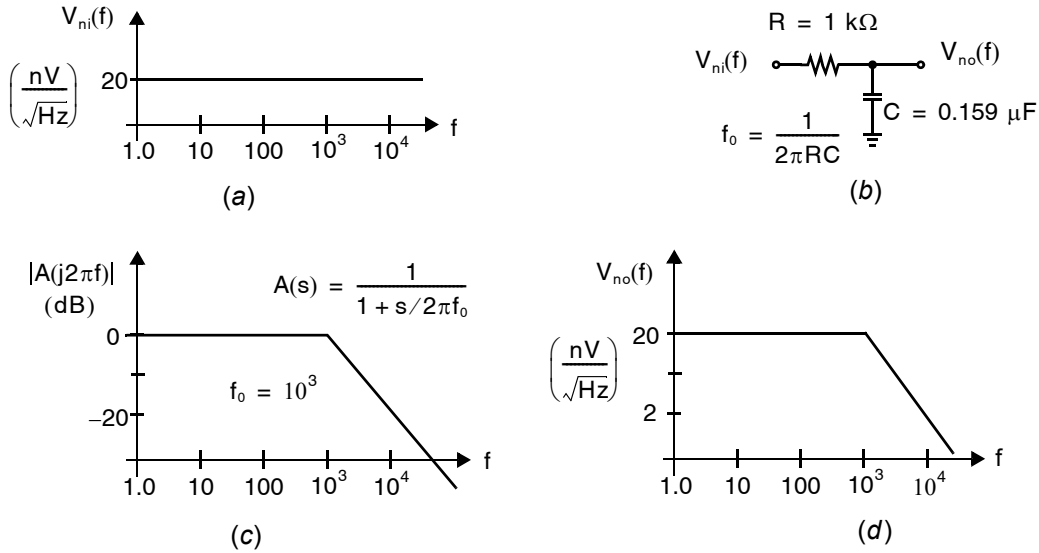


Fig. 9.8 (*a*) Spectral density for $V_{ni}(f)$. (*b*) RC filter to shape noise. (*c*) RC filter frequency response. (*d*) Spectral density for $V_{no}(f)$.

where $f_0 = 10^3$. Thus, the noise mean-squared value of $V_{no}(f)$ between dc and 100 kHz is given by

$$V_{no(rms)}^2 = \int_0^{10^5} \frac{20^2}{1 + \left(\frac{f}{f_0}\right)^2} df = 20^2 f_0 \arctan(f/f_0)\Big|_0^{10^5}$$
$$= 6.24 \times 10^5 (nV)^2 \tag{9.32}$$

which results in an rms value of $V_{no(rms)} = 0.79\ \mu V$ rms. Note that the noise rms value of $V_{no}(f)$ is almost $1/10$ that of $V_{ni}(f)$ since high-frequency noise above 1 kHz was filtered. The lesson here is that you should not design circuits for larger bandwidths than your signal requires, otherwise noise performance suffers.

9.2.5 Noise Bandwidth

> ***Key Point:*** *The noise bandwidth of a given filter is equal to the frequency span of a brick-wall filter that has the same rms output noise as the given filter has when white noise is applied to both filters, assuming the same peak gain for both filters.*

We just saw that the spectral density function is determined by the noise power within each 1-Hz bandwidth. Thus, in theory, one could measure the spectral density function by filtering a noise signal with a brick-wall bandpass filter having a 1-Hz bandwidth. The term *brick wall* implies here that the 1-Hz bandwidth of the filter is passed with a gain of one, whereas all other frequencies are entirely eliminated. However, practical filters can only approach a brick-wall response as their complexity (i.e. filter order) is increased. Thus, for lower-order filters with a 1-Hz passband, more noise power is passed than what is simply in a 1-Hz bandwidth. To account for the fact that practical filters have more gradual stopband characteristics, the term *noise bandwidth* is defined. *The noise bandwidth of a given filter is equal to the frequency span of a brick-wall filter that has the same rms output noise as the given filter has when white noise is applied to both filters, assuming the same peak gain for*

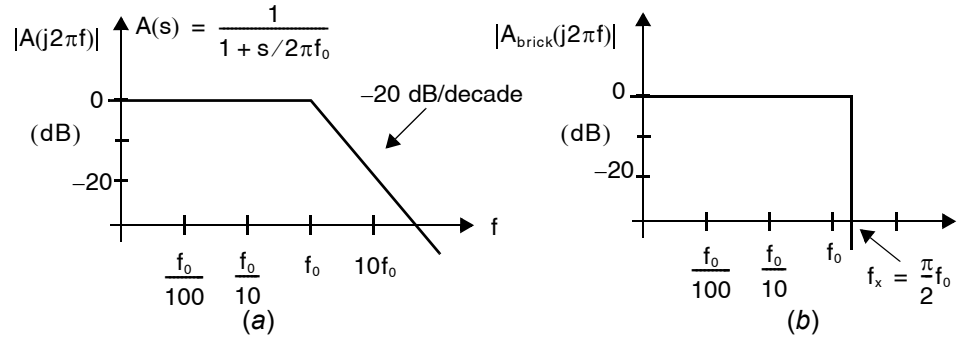


Fig. 9.9 (*a*) A first-order, low-pass response, and (*b*) a brick-wall filter that has the same peak gain and area as the first-order response.

both filters. In other words, given a filter response with peak gain A_0, the noise bandwidth is the width of a rectangular filter that has the same area and peak gain, A_0, as the original filter.

For example, consider a first-order, low-pass response with a 3-dB bandwidth of f_0, as shown in Fig. 9.9(*a*). Such a response would occur from the RC filter shown in Fig. 9.8(*b*) with $f_0 = (1/2\pi RC)$. The transfer function of $A(s)$ is given by

$$A(s) = \frac{1}{1 + \dfrac{s}{2\pi f_0}} \tag{9.33}$$

This results in the magnitude response of $A(s)$ being equal to

$$|A(j2\pi f)| = \left(\frac{1}{1 + \left(\dfrac{f}{f_0}\right)^2} \right)^{1/2} \tag{9.34}$$

An input signal, $V_{ni}(f)$, is a white noise source given by

$$V_{ni}(f) = V_{nw} \tag{9.35}$$

where V_{nw} is a constant. The total output noise rms value of $V_{no}(f)$ is equal to

$$V_{no(rms)}^2 = \int_0^\infty \frac{V_{nw}^2}{1 + \left(\dfrac{f}{f_0}\right)^2}\, df = V_{nw}^2 f_0 \arctan\left(\frac{f}{f_0}\right)\Bigg|_0^\infty = \frac{V_{nw}^2 \pi f_0}{2} \tag{9.36}$$

If this same input signal, $V_{ni}(f)$, is applied to the filter shown in Fig. 9.9(*b*), then the total output noise rms value equals,

$$V_{brick(rms)}^2 = \int_0^{f_x} V_{nw}^2\, df = V_{nw}^2 f_x \tag{9.37}$$

Finally, equating the two output noise rms values, $V_{no(rms)} = V_{brick(rms)}$, results in

$$f_x = \frac{\pi f_0}{2} \tag{9.38}$$

Thus, the noise bandwidth of a first-order, low-pass filter with a 3-dB bandwidth of f_0 equals $\pi(f_0/2)$. Note that, for the common case in which a first-order circuit is realized by a capacitor, C, and the resistance seen by that capacitor, R_{eq}, then

> ***Key Point:*** *The noise bandwidth of a first-order lowpass filter is $\pi/2$ times the filter's –3dB bandwidth.*

$$f_0 = \frac{1}{2\pi R_{eq} C} \tag{9.39}$$

and the noise bandwidth is given by

$$f_x = \frac{1}{4R_{eq}C} \tag{9.40}$$

The advantage of knowing the noise bandwidth of a filter is that, when white noise is applied to the filter input, the total output noise mean-squared value is easily calculated by multiplying the spectral density by the noise bandwidth. Specifically, in the first-order case just described, the total output noise mean-squared value, $V^2_{no(rms)}$, is equal to

$$V^2_{no(rms)} = V^2_{nw} f_x = V^2_{nw}\left(\frac{\pi}{2}\right) f_0 \tag{9.41}$$

Similar results for noise-bandwidth relationships can be obtained for higher-order and bandpass filters.

9.2.6 Piecewise Integration of Noise

Although simulation and computer techniques can analyze circuits or systems quite precisely (depending on their modeling accuracy), it is often convenient to make approximations that are useful in the early design stages. Such approximations are particularly useful in noise analysis, where large inaccuracies occur naturally due to such things as unknown parasitic effects and incomplete noise models. One such approximation is the estimation of total noise by integrating over frequency with the assumption that piecewise-linear Bode diagrams are exact. With such an approximation, integration formulas become much simpler when one needs only to integrate under linear functions and add together the resulting portions. The following example demonstrates this approach.

EXAMPLE 9.6

Consider an input noise signal, $V_{ni}(f)$, being applied to the amplifier $A(s)$, as shown in Fig. 9.10. Find the output noise rms value of V_{no} considering only frequencies above 1 Hz.

Solution

As shown, the output root spectral density, $V_{no}(f)$, is determined by the addition of the Bode diagrams for $V_{ni}(f)$ and $A(s)$. To perform piecewise integration of $V_{no}(f)$, the frequency range is broken into four regions, N_1 through N_4, as shown.

For the region N_1, the total mean-square noise is given by

$$N_1^2 = \int_1^{100} \frac{200^2}{f}\, df = 200^2 \ln(f)\Big|_1^{100} = 1.84 \times 10^5 (nV)^2 \tag{9.42}$$

In the region N_2, we have

$$N_2^2 = \int_{100}^{10^3} 20^2\, df = 20^2 f\Big|_{100}^{10^3} = 3.6 \times 10^5 (nV)^2 \tag{9.43}$$

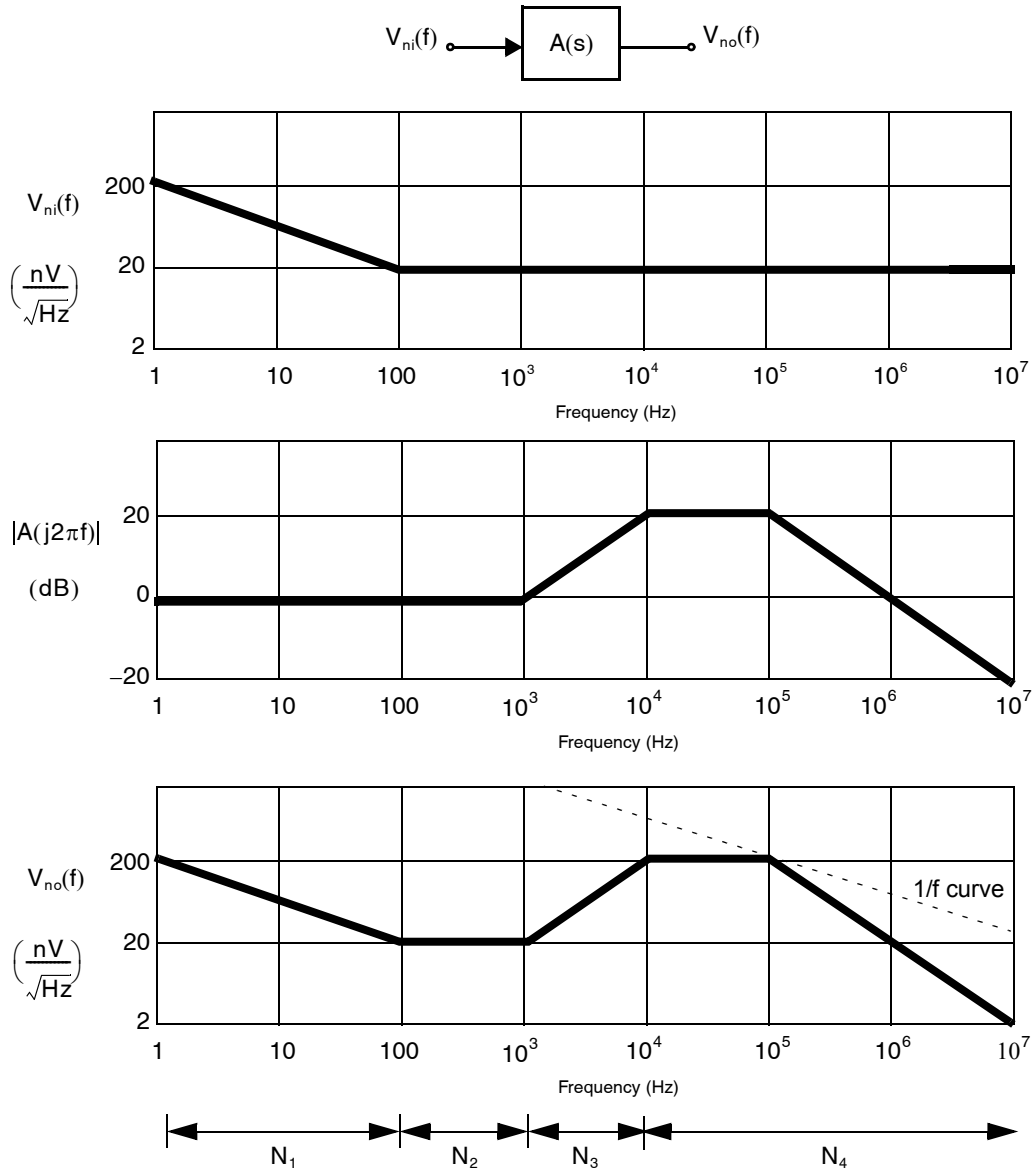


Fig. 9.10 Root spectral densities and amplifier curve example. $V_{no}(f)$ is the output noise that results from applying an input signal with noise $V_{ni}(f)$ to the amplifier $A(s)$.

Region N_3 ramps up rather than down, resulting in

$$N_3^2 = \int_{10^3}^{10^4} \frac{20^2 f^2}{(10^3)^2}\, df = \left(\frac{20}{10^3}\right)^2 \left[\frac{1}{3} f^3 \Big|_{10^3}^{10^4}\right] = 1.33 \times 10^8 (nV)^2 \tag{9.44}$$

Finally, for region N_4, we can use the noise bandwidth result of a first-order, low-pass response and simply remove the noise portion resulting from under 10^4 Hz. Specifically, we have

$$N_4^2 = \int_{10^4}^{\infty} \frac{200^2}{1+\left(\frac{f}{10^5}\right)^2}\,df = \int_0^{\infty} \frac{200^2}{1+\left(\frac{f}{10^5}\right)^2}\,df - \int_0^{10^4} 200^2\,df$$

$$= 200^2\left(\frac{\pi}{2}\right)10^5 - (200^2)(10^4) = 5.88\times 10^9 (\text{nV})^2 \tag{9.45}$$

Thus, the total output noise can be estimated to be

$$V_{no(rms)} = (N_1^2 + N_2^2 + N_3^2 + N_4^2)^{1/2} = 77.5\ \mu\text{V rms} \tag{9.46}$$

An interesting point to note here is that in the preceding example, $N_4 = 76.7\ \mu$V rms is quite close to the total noise value of 77.5 μV rms. Thus, in practice, there is little need to find the noise contributions in the regions N_1, N_2, and N_3. Such an observation leads us to the 1/f noise tangent principle.

9.2.7 1/f Noise Tangent Principle

The *1/f noise tangent principle* is as follows: To determine the frequency region or regions that contribute the dominant noise, *lower a 1/f noise line until it touches the spectral density curve—the total noise can be approximated by the noise in the vicinity of the 1/f line* [Kennedy, 1988; Franco, 2002]. For example, lowering a 1/f line toward the root spectral density of $V_{no}(f)$ in Fig. 9.10 indicates that the noise around 10^5 dominates. The reason this simple rule works is that a curve proportional to $1/x$ results in equal power over each decade of frequency. Therefore, by lowering this constant power/frequency curve, the largest power contribution will touch it first. However, because the integration of $1/x$ approaches infinity if either the upper bound is infinity or the lower bound is zero, one must be careful in cases where the spectral density curve runs parallel to a 1/f tangent line for an appreciable frequency span. For example, consider once again Fig. 9.10, where the region N_1 runs parallel to the 1/f tangent line. However, in this example, region N_1 does not contribute much noise since the noise was only integrated above 1 Hz. If a much lower frequency bound is used, this region can also contribute appreciable noise power.

> ***Key Point:*** *To determine the frequency region or regions that contribute the dominant noise, lower a 1/f noise line until it touches the spectral density curve. The total noise can be approximated by the noise in the vicinity of the 1/f line.*

9.3 NOISE MODELS FOR CIRCUIT ELEMENTS

There are three main fundamental noise mechanisms: thermal, shot, and flicker. In this section, we discuss noise models for popular circuit elements where all three mechanisms occur. However, first we briefly describe these noise phenomena.

Thermal noise is due to the thermal excitation of charge carriers in a conductor. This noise has a white spectral density and is proportional to absolute temperature. It is not dependent on bias conditions (dc bias current) and it occurs in all resistors (including semiconductors) above absolute zero temperature. Thus, thermal noise places fundamental limits on the dynamic range achievable in electronic circuits. It should be mentioned here that thermal noise is also referred to as Johnson or Nyquist noise since it was first observed by J. B. Johnson [Johnson, 1928] and analyzed using the second law of thermodynamics by H. Nyquist [Nyquist, 1928].

Shot noise was first studied by W. Schottky using vacuum-tube diodes [Schottky, 1918], but shot noise also occurs in pn junctions. This noise occurs because the dc bias current is not continuous and smooth but instead is a result of pulses of current caused by the flow of individual carriers. As such, *shot noise is dependent on the dc bias current*. It can also be modelled as a white noise source. Shot noise is also typically larger than thermal noise and is sometimes used to create white noise generators.

Flicker noise is the least understood of the three noise phenomena. It is found in all active devices as well as in carbon resistors,[4] but it occurs only when a dc current is flowing. Flicker noise usually arises due to traps in the semiconductor, where carriers that would normally constitute dc current flow are held for some time period and then released. Flicker noise is also commonly referred to as 1/f noise since it is well modelled as having a $1/f^{\alpha}$ spectral density, where α is between 0.8 and 1.3. Although both bipolar and MOSFET transistors have flicker noise, it is a significant noise source in MOS transistors, whereas it can often be ignored in bipolar transistors.

9.3.1 Resistors

The major source of noise in resistors is thermal noise. As just discussed, it appears as white noise and can be modelled as a voltage source, $V_R(f)$, in series with a noiseless resistor. With such an approach, the spectral density function, $V_R^2(f)$, is found to be given by

$$V_R^2(f) = 4kTR \tag{9.47}$$

where k is Boltzmann's constant (1.38×10^{-23} JK^{-1}), T is the temperature in Kelvins, and R is the resistance value.

Alternatively, (9.47) may be rewritten noting that a 1-kΩ resistor exhibits a root spectral density of 4.06 $nV/\sqrt{Hz}$ in thermal noise at room temperature (300 °K). Since the root spectral density is proportional to the square root of the resistance, we can also write

$$V_R(f) = \sqrt{\frac{R}{1k}} \times 4.06 \text{ nV}/\sqrt{\text{Hz}} \quad \text{for } 27\ ^\circ\text{C} \tag{9.48}$$

> ***Key Point:*** *Resistors exhibit thermal noise modeled either by a voltage source in series with the resistor having a white noise spectral density* $4kTR$*, or a Norton equivalent current source having a white noise spectral density* $4kT/R$.

Note that, to reduce the thermal noise voltage due to resistors, one must either lower the temperature or use lower resistance values. The fact that lower resistance values cause less thermal noise becomes much more apparent when we look at kT/C noise in capacitors later in this section.

An alternate model can be derived by finding the Norton equivalent circuit. Specifically, the series voltage noise source, $V_R(f)$, can be replaced with a parallel current noise source, $I_R(f)$, given by

$$I_R^2(f) = \frac{V_R^2(f)}{R^2} = \frac{4kT}{R} \tag{9.49}$$

Both resistor models are summarized in Fig. 9.11.

9.3.2 Diodes

> ***Key Point:*** *Diodes exhibit shot noise modeled by a current source having a white noise spectral density* $2qI_D$.

Shot noise is typically the dominant noise in diodes and can be modelled with a current source in parallel with the small-signal resistance of the diode, as Fig. 9.11 shows. The spectral density function of the current source is found to be given by

$$I_d^2(f) = 2qI_D \tag{9.50}$$

4. Carbon resistors are not used in integrated-circuit design but are available as discrete elements.

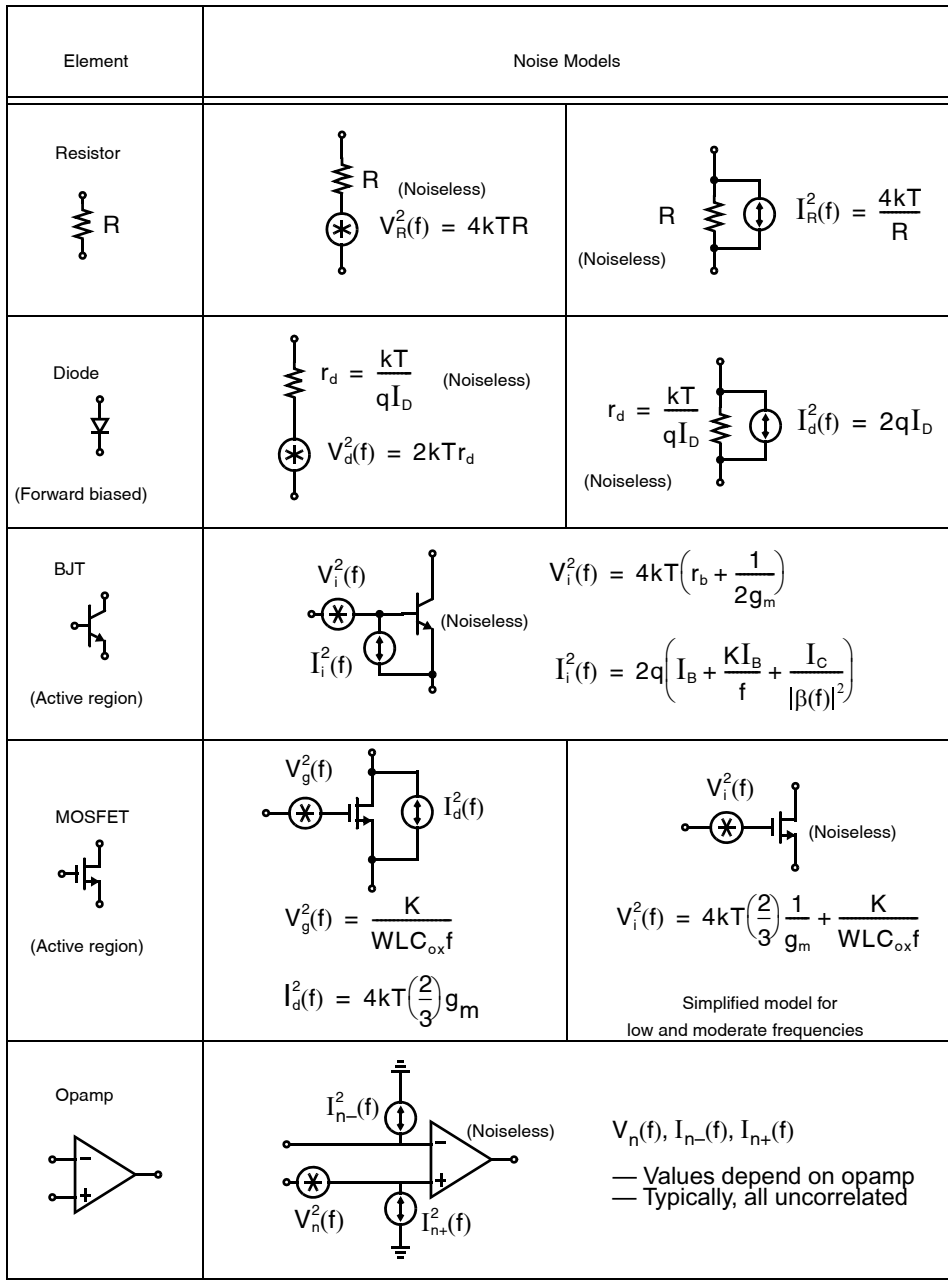


Fig. 9.11 Circuit elements and their noise models. Note that capacitors and inductors do not generate noise.

where q is one electronic charge $(1.6 \times 10^{-19}\ \mathrm{C})$ and I_D is the dc bias current flowing through the diode. The small-signal resistance of the diode, r_d, is given by the usual relationship,

$$r_d = \frac{kT}{qI_D} \tag{9.51}$$

The Thévenin equivalent circuit can also be used, as shown in Fig. 9.11. Note that the small-signal resistance, r_d, is used for small-signal modelling and is not a physical resistor; hence, r_d does not contribute any thermal noise.

9.3.3 Bipolar Transistors

The noise in bipolar transistors is due to the shot noise of both the collector and base currents, the flicker noise of the base current, and the thermal noise of the base resistance. A common practice is to combine all these noise sources into two equivalent noise sources at the base of the transistor, as shown in Fig. 9.11. Here, the equivalent input voltage noise, $V_i(f)$, is given by

$$V_i^2(f) = 4kT\left(r_b + \frac{1}{2g_m}\right) \tag{9.52}$$

where the r_b term is due to the thermal noise of the base resistance and the g_m term is due to collector-current shot noise referred back to the input. The equivalent input current noise, $I_i(f)$, equals

$$I_i^2(f) = 2q\left(I_B + \frac{KI_B}{f} + \frac{I_C}{|\beta(f)|^2}\right) \tag{9.53}$$

Key Point: *Bipolar transistor noise is typically dominated by the thermal noise of the series base resistance, and shot noise in the base-current junction.*

where the $2qI_B$ term is a result of base-current shot noise, the KI_B/f term models 1/f noise (K is a constant dependent on device properties), and the I_C term is the input-referred collector-current shot noise (often ignored).

The noise of r_b typically dominates in $V_i(f)$, and the base-current shot noise often dominates in the input-current noise, $I_i^2(f)$. Thus, the equivalent voltage and current noise are not derived from the same noise sources. As a result, it is common practice to assume that the input voltage and current noise sources in a bipolar transistor are uncorrelated.[5]

9.3.4 MOSFETS

The dominant noise sources for active MOSFET transistors are flicker and thermal noise, as shown in Fig. 9.11. The flicker noise is modelled as a voltage source in series with the gate of value

$$V_g^2(f) = \frac{K}{WLC_{ox}f} \tag{9.54}$$

Key Point: *At low frequencies, 1/f noise is dominant in MOSFET circuits. It is minimized by using large gate areas and p-channel transistors wherever possible.*

where the constant K is dependent on device characteristics and can vary widely for different devices in the same process. The variables W, L, and C_{ox} represent the transistor's width, length, and gate capacitance per unit area, respectively. *The 1/f noise is inversely proportional to the transistor area,* WL, so larger devices have less 1/f noise. In MOSFET circuits, 1/f noise is extremely important because it typically dominates at low frequencies unless switching-circuit techniques are used to reduce its effect. Typically, the 1/f noise constant K is

5. An exception to this is high-frequency designs in which the collector shot noise becomes more important because, at high frequencies, β becomes small. In this case, the input current noise source is partially correlated with the input voltage noise source. If neither noise source dominates, then the correct analysis is more difficult and beyond the scope of this text. Fortunately, this case is not often encountered in practice.

smaller for p-channel transistors than their n-channel counterparts since their majority carriers (holes) are less likely to be trapped, so it is desirable to use p-channel transistors when 1/f noise is to be reduced.

The derivation of the thermal noise term is straightforward and is due to the resistive channel of a MOS transistor in the active region. If the transistor was in triode, the thermal noise current in the drain due to the channel resistance would simply be given by $I_d^2(f) = (4kT)/r_{ds}$, where r_{ds} is the channel resistance. However, when the transistor is in the active region, the channel cannot be considered homogeneous, and thus, the total noise is found by integrating over small portions of the channel. Such an integration results in the noise current in the drain given by

$$I_d^2(f) = 4kT\gamma g_m \tag{9.55}$$

For the case $V_{DS} = V_{GS} - V_T$ and assuming a long channel device, $\gamma = 2/3$. However, for short gate-length devices much higher values of γ may be observed. Note that the white noise parameter γ is different from the body-effect parameter γ.

Key Point: *MOSFET thermal noise is simply that of a resistor with value* r_{ds} *when in triode. In active mode, thermal noise is modeled by a white current noise source between drain and source with spectral density* $8kTg_m/3$*, for a long channel device, and somewhat higher for short-channel devices.*

Often, noise analyses are done just by including this noise source between the transistor drain and source. Sometimes, however, analysis may be simplified if it is replaced by an equivalent input noise source. To find the equivalent noise voltage that would cause this drain current, we note that the drain current is equal to the gate voltage times the transconductance of the device, or, mathematically, $I_d(f) = g_m V_i(f)$. Thus, dividing (9.55) by g_m^2 results in the simplified MOSFET model, also shown in Fig. 9.11, where there is now only one voltage noise source. However, one should be aware that this simplified model assumes the gate current is zero. Although this assumption is valid at low and moderate frequencies, an appreciable amount of current would flow through the gate-source capacitance, C_{gs}, at higher frequencies. In summary, although most noise analysis can be performed using the simplified model, if in doubt, one should use the model with the thermal noise placed as a current source in parallel with the drain-source channel.

Although gate leakage can introduce an additional source of noise, no gate leakage noise terms have been included in this noise model since the gate leakage is generally so small that its noise contribution is rarely significant.

EXAMPLE 9.7

A large MOS transistor consists of ten individual transistors connected in parallel. Considering 1/f noise only, what is the equivalent input voltage noise spectral density for the ten transistors compared to that of a single transistor?

Solution

From (9.54), the 1/f noise can be modelled as a current source going from the drain to the source with a noise spectral density of

$$I_d^2(f) = \frac{Kg_m^2}{WLC_{ox}f} \tag{9.56}$$

where g_m is the transconductance of a single transistor. When ten transistors are connected in parallel, the drain-current noise spectral density is ten times larger, so we have

$$I_{d=10}^2(f) = \frac{10Kg_m^2}{WLC_{ox}f} \tag{9.57}$$

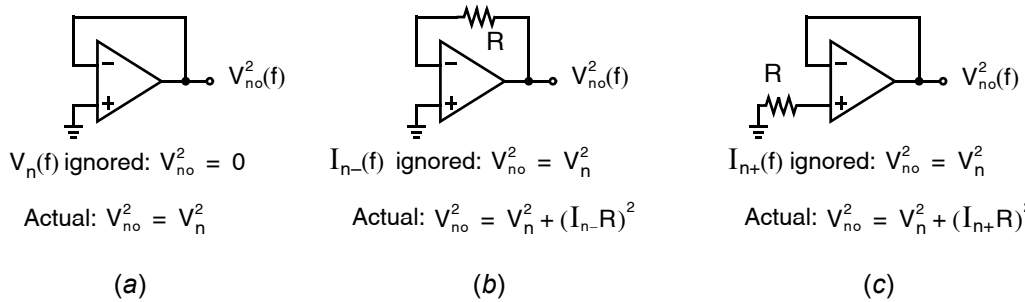


Fig. 9.12 Opamp circuits showing the need for three noise sources in an opamp noise model. Assume the resistance, R, is noiseless. Also, notation is simplified from $V_n(f)$ to V_n, and so on.

When this noise is referred back to the input of the equivalent transistor, we have

$$V_{i=10}^2(f) = \frac{10Kg_m^2}{WLC_{ox}fg_{m=10}^2} = \frac{K}{10WLC_{ox}f} \tag{9.58}$$

since the transconductance of ten transistors in parallel, $g_{m=10}$, is equal to 10 g_m. Thus, the drain current noise spectral density of the equivalent transistor is ten times larger, but the input voltage noise spectral density is ten times smaller. This result is expected because the input voltage noise source due to 1/f noise is inversely proportional to the equivalent transistor area, which in this case is 10 WL.

9.3.5 Opamps

Noise in opamps is modelled using three uncorrelated input-referred noise sources, as shown in Fig. 9.11. With an opamp that has a MOSFET input stage, the current noises can often be ignored at low frequencies since their values are small. However, for bipolar input stages, all three noise sources are typically required, as shown in Fig. 9.12. In Fig. 9.12(*a*), if $V_n(f)$ is not included in the model, a unity-gain buffer with no resistors indicates that the circuit output is noiseless. In fact, the voltage noise source, $V_n(f)$, dominates. If $I_{n-}(f)$ is not included in an opamp model, the circuit shown in Fig. 9.12(*b*) indicates that the output noise voltage equals $V_n(f)$. Here, the current noise may dominate if the resistance, R, is large. A similar conclusion can be drawn with $I_{n+}(f)$, as shown in Fig. 9.12(*c*).

9.3.6 Capacitors and Inductors

Capacitors and inductors do not generate any noise. However, they do accumulate noise generated by other noise sources. Here, we will see that the capacitor noise mean-squared voltage equals kT/C when it is connected to an arbitrary resistor value.

Consider a capacitance, C, in parallel with a resistor of arbitrary size, R, as shown in Fig. 9.13(*a*). The equivalent circuit for noise analysis is shown in Fig. 9.13(*b*). To determine the total noise mean-squared value across the capacitor, we note that $V_{no}(f)$ is simply a first-order, low-pass, filtered signal with $V_R(f)$ as the input. Therefore, we recognize that the noise bandwidth is given by $(\pi/2)f_0$ as in (9.41), and since the input has a white spectral density, the total output mean-squared value is calculated as

$$\begin{aligned} V_{no(rms)}^2 &= V_R^2(f)\left(\frac{\pi}{2}\right)f_0 = (4kTR)\left(\frac{\pi}{2}\right)\left(\frac{1}{2\pi RC}\right) \\ V_{no(rms)}^2 &= \frac{kT}{C} \end{aligned} \tag{9.59}$$

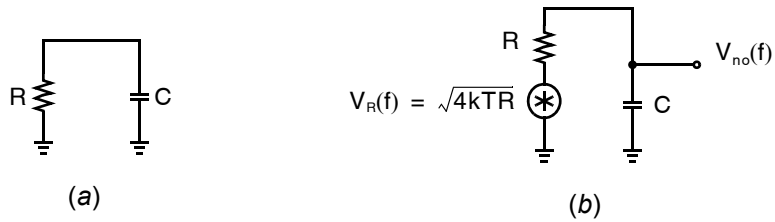


Fig. 9.13 (*a*) Capacitor, C, in parallel with a resistor, and (*b*) equivalent noise model circuit.

In other words, the rms voltage value across a capacitor is equal to $\sqrt{kT/C}$, whenever a simple passive resistance of any value is connected across it. Such a result is due to fact that small resistances have less noise spectral density but result in a wide bandwidth, compared to large resistances, which have reduced bandwidth but larger noise spectral density.

> ***Key Point:*** *In a circuit comprising only capacitors and resistors (including triode MOSFETs) but no active devices, the squared rms noise on each capacitor is* kT/C.

Finally, it should be stated that this noise property for capacitors gives a fundamental limit on the minimum noise level across a capacitor connected to passive components.[6] Thus, to lower the noise level either the temperature must be lowered, or the capacitance value must be increased in which case a constant bandwidth is maintained by decreasing resistance.

EXAMPLE 9.8

At a room temperature of 300 °K , what capacitor size is needed to achieve a 96-dB dynamic range in an analog circuit with maximum signal levels of 1 V rms?

Solution

The value of noise that can be tolerated here is 96 dB down from 1 V rms, which is

$$V_{n(rms)} = \frac{1V}{10^{96/20}} = 15.8\ \mu V \text{ rms} \tag{9.60}$$

Using (9.59), we have

$$C = \frac{kT}{V_{n(rms)}^2} = 16.6 \text{ pF} \tag{9.61}$$

Thus, the minimum capacitor size that can be used (without oversampling) is 16.6 pF. The use of this minimum capacitor size determines the maximum resistance size allowed to achieve a given time constant.

Finally, it should be mentioned that the equivalent noise current mean-squared value in an inductor of value L connected only to resistors is given by (see Problem 9.19)

$$I_{no(rms)}^2 = \frac{kT}{L} \tag{9.62}$$

6. Some feedback circuits can make the noise smaller but signal levels are also typically smaller (see Problem 9.18).

9.3.7 Sampled Signal Noise

In many cases, one should obtain a sampled value of an analog voltage. For example, switched-capacitor circuits make extensive use of sampling, and sample-and-hold circuits are commonly used in analog-to-digital and digital-to-analog conversion.

Key Point: *Sampling an analog voltage on a capacitor captures both the signal component and noise with a mean-square value of* kT/C*, assuming a purely passive sampling network.*

Consider a basic sample-and-hold circuit, as shown in Fig. 9.14. When ϕ_{clk} drops, the transistor turns off and, in an ideal noiseless world, the input voltage signal at that instance would be held on capacitance C. However, when thermal noise is present, the resistance when the transistor is switched on causes voltage noise on the capacitor with an rms value of $\sqrt{kT/C}$. Thus, when the switch is turned off, the noise as well as the desired signal is held on C. As a result, a fundamental limit occurs for sampled signals using a capacitance C—an rms noise voltage of $\sqrt{kT/C}$. The sampled noise may be higher than this limit if the sample-and-hold circuit includes active components, such as an opamp.

The sampled noise voltage does not depend on the sampling rate and is independent from sample to sample. This fact suggests a method to reduce the noise level of a signal measurement. Specifically, in the case where v_{in} is a dc (or low-frequency) signal, taking only one sample results in a noise voltage of $\sqrt{kT/C}$. However, if many samples are taken (say, 1,000) and all samples are averaged, the averaged value will have a reduced noise level. The reason this technique improves the measurement accuracy is that, when individual sampled values are summed together, their signal values add linearly, whereas their noise values (being uncorrelated) add as the root of the sum of squares. This technique is known as oversampling and will be discussed at length with respect to oversampling converters in Chapter 18.

9.3.8 Input-Referred Noise

Key Point: *The input-referred noise of a circuit, if applied to the input of a noiseless copy of the circuit, results in the exact same output noise as when all of the circuit's noise sources were present. It is found by dividing the observed output noise by the gain of the circuit.*

In general, the noise voltage at a particular node, or noise current in a particular branch, will be a superposition of multiple filtered uncorrelated noise sources, as illustrated in Fig. 9.7. In order to quantify the impact of all these noise sources upon SNR, it is useful to know the total input-referred noise of the circuit.

The input-referred noise of a circuit, if applied to the input of a noiseless copy of the circuit, results in the exact same output noise as when all of the circuit's noise sources were present. It is found by dividing the observed output noise by the gain of the circuit. A method to determine input referred noise by analysis or simulation is illustrated in Fig. 9.15. The output noise of the circuit $v_{on(rms)}$ is first determined, then divided by the circuit's mid-band gain to arrive at a rms input-referred noise quantity. If the input source is best modeled as a voltage source, it makes sense to use an input-referred noise voltage defined in terms of the circuit's mid-band small signal voltage gain,

$$v_{in(rms)} = v_{on(rms)}/A \tag{9.63}$$

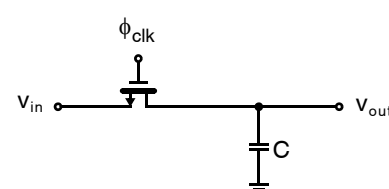


Fig. 9.14 A sample-and-hold circuit.

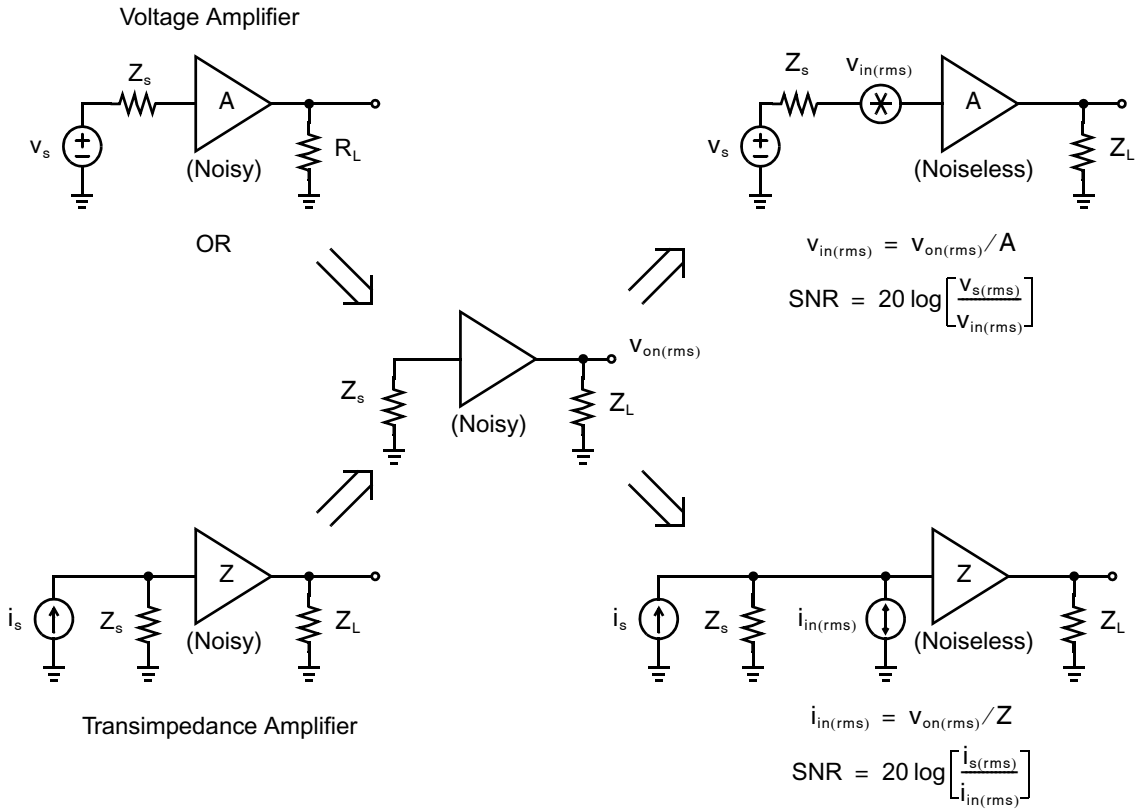


Fig. 9.15 Determination of input-referred noise.

Alternatively, if the input source is best modeled as a Norton-equivalent current source, the input-referred noise current may be a more relevant quantity, obtained from the circuit's mid-band transimpedance gain,

$$i_{in(rms)} = v_{on(rms)}/Z \tag{9.64}$$

In general, the output noise depends upon the source and load impedances connected to the circuit, Z_s and Z_L in Fig. 9.15, so noise analysis should always be performed with those impedances in place. However, when it is known that Z_s is much smaller than the circuit's input impedance (i.e., a voltage input) then the noise analysis can be performed with $Z_s = 0$ (short-circuit) with good accuracy. For example, this is typically the case for an opamp. Similarly, if Z_s is much larger than the circuit input impedance (i.e., current input) then the noise analysis can be performed with an open-circuit input.

EXAMPLE 9.9

Two voltage amplifiers (each having very large input impedance and small output impedance) are available: one with a gain of 3 V/V and 3 μV_{rms} noise observed at the output; the other with a gain of 8 V/V and 6 μV_{rms} noise observed at its output. What is the input-referred noise of each amplifier? If the two amplifiers are to be placed in series to realize a gain of 24 V/V, in what order should they be placed in order to obtain the best noise performance? What is the resulting input-referred noise of the overall system?

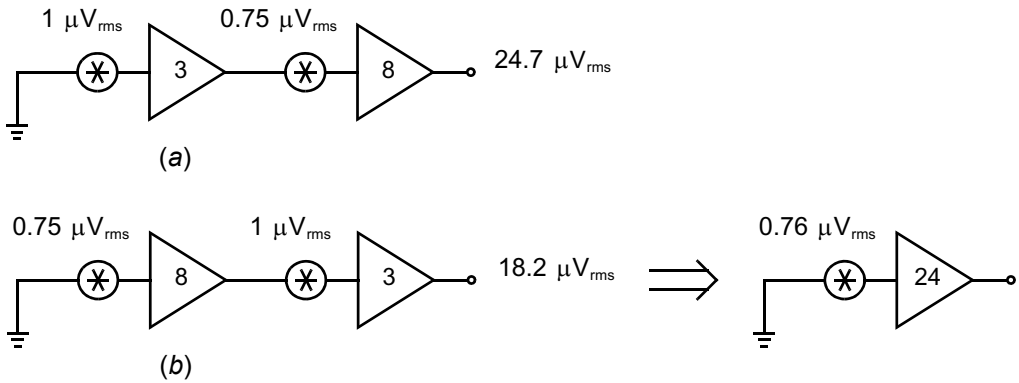


Fig. 9.16 Computing the input-referred noise of Example 9.9.

Solution

The first amplifier has an input referred noise of $3\ \mu V/3 = 1\ \mu V_{rms}$ and the second has an input-referred noise of $6\ \mu V/8 = 0.75\ \mu V_{rms}$. First consider placing the amplifier with a gain of 3 V/V first, as illustrated in Fig. 9.16(*a*). In this case, the total output noise is given by

$$v_{on(rms)}^2 = (1\ \mu V)^2(3 \cdot 8)^2 + (0.75\ \mu V)^2 8^2 = 0.61 \cdot 10^{-9}\ V^2$$

$$v_{on(rms)} = 24.7\ \mu V_{rms}$$

The alternative arrangement is depicted in Fig. 9.16(*b*), and results in an output noise of

$$v_{on(rms)}^2 = (0.75\ \mu V)^2(8 \cdot 3)^2 + (1\ \mu V)^2 3^2 = 0.33 \cdot 10^{-9}\ V^2$$

$$v_{on(rms)} = 18.2\ \mu V_{rms}$$

Clearly, the second situation is preferable. The total input referred noise in this case is

$$v_{in(rms)} = 18.2\ \mu V_{rms}/(3 \cdot 8) = 0.76\ \mu V_{rms}$$

Notice that the total input referred noise is almost identical to that of the first amplifier alone. This result could also have been obtained by input-referring the noise contribution of each amplifier directly from Fig. 9.16(*b*).

$$v_{in(rms)} = \sqrt{(0.75\ \mu V)^2 + \frac{(1\ \mu V)^2}{8^2}} = 0.76\ \mu V_{rms}$$

The larger gain in the first stage renders the noise from the second stage negligible. This illustrates why, for example, the first stage is the dominant noise contributor in a two-stage opamp. In general, when very low-amplitude signals must be processed with high fidelity, it is desirable to use an initial low-noise and high-gain pre-amplifier. Doing so ensures the noise performance of subsequent stages is not critical.

The noise spectral density at the output of the circuit is generally the sum of noise contributed by the circuit itself, $v_{ao}(f)$, as well as thermal noise from the real part of the source impedance filtered by the circuit, $4kTR_s|A(f)|^2$. These quantities may be input referred as illustrated in Fig. 9.17 for a voltage amplifier where

$$v_{ai}^2(f) = \frac{v_{ao}^2(f)}{|A(f)|^2} \tag{9.65}$$

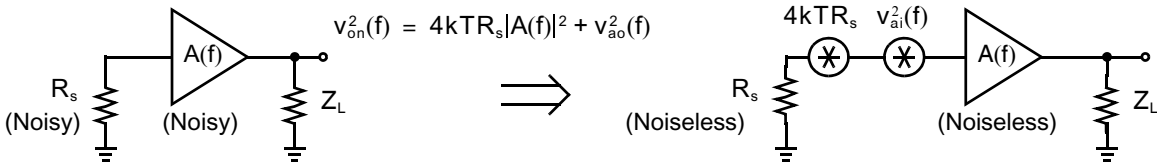


Fig. 9.17 Separating noise contributed by the source resistance and circuit noise.

Note that the input-referred noise $v_{in(rms)}$ in (9.63) is not obtained by integrating the input-referred noise spectrum over some bandwidth. The rms input-referred noise divides the rms output noise by the mid-band gain only so that it can be used to determine SNR assuming the input signal is confined to the circuit's mid-band region.

Key Point: *The noise performance of a circuit at a particular frequency is usually reported as its input-referred noise spectrum, equal to the circuit's output noise spectrum divided by its magnitude response, or equivalently by its noise figure.*

In radio-frequency applications, the signal input is narrowband and filters can be used to eliminate all out-of-band noise. In this case, a measure of noise spectral density at a particular frequency, also called the *spot noise*, is useful. A common noise metric in this case is the *noise factor*, $F(f)$, given by the ratio of the total noise to the noise that would arise if only the source resistance were noisy, at a particular frequency.

$$F(f) = \frac{v_{no}^2(f)}{4kTR_s|A(f)|^2} \tag{9.66}$$

This quantity may be related to the input referred noise of the amplifier alone, $v_{ai}^2(f)$.

$$F(f) = \frac{4kTR_s + v_{ai}^2(f)}{4kTR_s} = 1 + \frac{v_{ai}^2(f)}{4kTR_s} \tag{9.67}$$

Expressed in decibels, the noise factor becomes the noise figure of the circuit,

$$NF(f) = 10\log_{10}[F(f)]\ \text{dB} \tag{9.68}$$

9.4 NOISE ANALYSIS EXAMPLES

In this section, a variety of circuits are analyzed from a noise perspective. Although some useful design techniques for reducing noise are presented, the main purpose of this section is to give the reader some experience in analyzing circuits from a noise perspective.

9.4.1 Opamp Example

Consider an inverting amplifier and its equivalent noise model, as shown in Fig. 9.18. Here, $V_n(f)$, $I_{n-}(f)$, and $I_{n+}(f)$ represent the opamp's equivalent input noise, and the remaining noise sources are resistor thermal noise sources. Note that current noise sources are used in the models for R_1 and R_f, whereas a voltage noise source is used for R_2. As we will see, these choices of noise sources simplify the circuit analysis.

First, using superposition and assuming all noise sources are uncorrelated, consider the output noise voltage, $V_{no1}^2(f)$, due only to I_{n1}, I_{nf}, and I_{n-}. These three noise currents add together, and their total current sum is fed into the parallel combination of C_f and R_f. Note that no current passes through R_1 since the voltage across it is zero due to the virtual ground at the negative opamp terminal (assuming a high-gain opamp). Thus, the output noise mean-squared value due to these three noise currents is given by

$$V_{no1}^2(f) = [I_{n1}^2(f) + I_{nf}^2(f) + I_{n-}^2(f)]\left|\frac{R_f}{1 + j2\pi f R_f C_f}\right|^2 \tag{9.69}$$

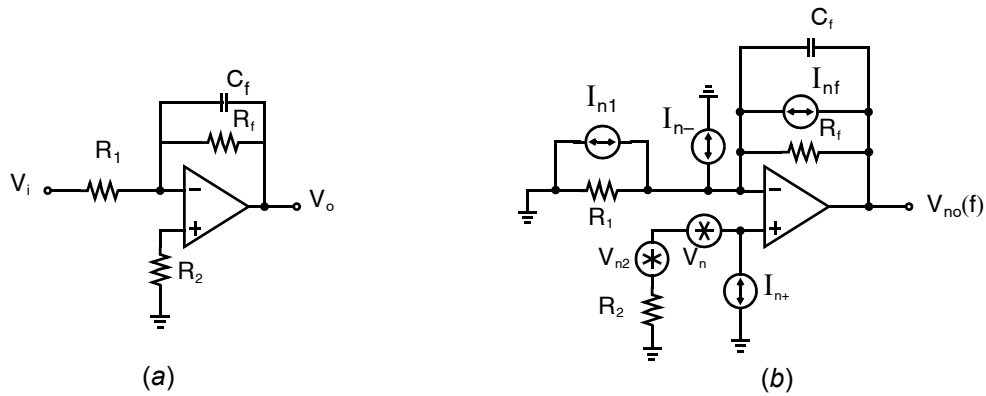


Fig. 9.18 (*a*) Low-pass filter, and (*b*) equivalent noise model.

This equation indicates that this part of the output noise value equals the sum of the noise currents multiplied by R_f^2. This noise portion is then shaped by a low-pass filter with a 3-dB frequency equal to $f_0 = 1/(2\pi R_f C_f)$ Following an analysis similar to that in Section 9.3.6, the total integrated mean squared output noise voltage due to Rf alone may be shown to be simply kT/C_f. Including the noise of R_1 increases the result to $(1 + R_f/R_1)kT/C_f$ and including opamp noise increases it further. Hence, kT/C_f is a lower limit on the noise of opamps with capacitive feedback.

Using superposition further, the output mean-squared value, $V_{no2}^2(f)$, due to the three noise sources at the positive opamp terminal can be found as follows: By converting I_{n+} to a voltage source (by multiplying it by R_2), we see that the three noise voltages are summed and are applied to the positive opamp terminal. Since the gain from the positive opamp terminal to the output signal is easily found, the output noise mean-squared value due to these three noise sources is given by

$$V_{no2}^2(f) = [I_{n+}^2(f)R_2^2 + V_{n2}^2(f) + V_n^2(f)]\left|1 + \frac{R_f/R_1}{1 + j2\pi f C_f R_f}\right|^2 \tag{9.70}$$

This equation indicates that this part of the output noise mean-squared value equals the sum of the noise voltages, and this noise portion is then shaped by the shown transfer function. For this transfer function, if $R_f << R_1$, then its gain approximately equals unity for all frequencies. Thus, for an ideal opamp, the noise would exist up to infinite frequency, resulting in an infinite amount of mean-squared volts. However, for practical circuits, the gain drops off above the unity-gain frequency of the opamp, and thus the noise is effectively low-pass filtered. In the case where $R_f >> R_1$, the low-frequency gain is roughly R_f/R_1, and its 3-dB frequency is the same as in the case of the noise sources at the negative opamp terminal (i.e., $f_0 = 1/(2\pi R_f C_f)$). However, in this case, the gain only decreases to unity and then remains at that level. Treating the Bode plot as exact and noting that, above f_0, the gain drops off at –20 dB/decade, this transfer function reaches unity around $f_1 = (R_f/R_1)f_0$. Thus, one should also include the opamp's positive input noise (with a gain of one) integrated over the region between f_1 and the unity-gain frequency of the opamp.

Finally, the total output noise mean-squared value is simply the sum

$$V_{no}^2(f) = V_{no1}^2(f) + V_{no2}^2(f) \tag{9.71}$$

or, if rms values are found,

$$V_{no(rms)}^2 = V_{no1(rms)}^2 + V_{no2(rms)}^2 \tag{9.72}$$

EXAMPLE 9.10

Estimate the total output noise rms value for a 10-kHz low-pass filter, as shown in Fig. 9.18, when $C_f = 160$ pF, $R_f = 100$ k, $R_1 = 10$ k, and $R_2 = 9.1$ k. Also, find the SNR for an input signal equal to 100 mV rms. Assume that the noise voltage of the opamp is given by $V_n(f) = 20\ \text{nV}/\sqrt{\text{Hz}}$, both its noise currents are $I_n(f) = 0.6\ \text{pA}/\sqrt{\text{Hz}}$, and that its unity-gain frequency equals 5 MHz.

Solution

Assuming the device is at room temperature, the resistor noise sources are equal to

$$I_{nf} = 0.406\ \text{pA}/\sqrt{\text{Hz}} \tag{9.73}$$

$$I_{n1} = 1.28\ \text{pA}/\sqrt{\text{Hz}} \tag{9.74}$$

$$V_{n2} = 12.2\ \text{nV}/\sqrt{\text{Hz}} \tag{9.75}$$

The low-frequency value of $V_{no1}^2(f)$ is found by letting $f = 0$ in (9.69).

$$\begin{aligned} V_{no1}^2(0) &= [I_{n1}^2(0) + I_{nf}^2(0) + I_{n-}^2(0)]R_f^2 \\ &= (0.406^2 + 1.28^2 + 0.6^2)(1 \times 10^{-12})^2(100\,\text{k})^2 \\ &= (147\ \text{nV}/\sqrt{\text{Hz}})^2 \end{aligned} \tag{9.76}$$

Since (9.69) also indicates that this noise is low-pass filtered, the rms output noise value due to these three sources can be found using the concept of a noise equivalent bandwidth. Specifically, we multiply the spectral density by $(\pi f_0)/2$, where $f_0 = 1/(2\pi R_f C_f)$. Thus,

$$\begin{aligned} V_{no1(rms)}^2 &= (147\ \text{nV}/\sqrt{\text{Hz}})^2 \times \frac{1}{4(100\ \text{k}\Omega)(160\ \text{pF})} \\ &= (18.4\ \mu\text{V})^2 \end{aligned} \tag{9.77}$$

To estimate the output noise due to the sources at the positive opamp terminal, we find $V_{no2}^2(0)$ to be given by

$$\begin{aligned} V_{no2}^2(0) &= [I_{n+}^2(f)R_2^2 + V_{n2}^2(f) + V_n^2(f)](1 + R_f/R_1)^2 \\ &= (24.1\ \text{nV}/\sqrt{\text{Hz}})^2 \times 11^2 \\ &= (265\ \text{nV}/\sqrt{\text{Hz}})^2 \end{aligned} \tag{9.78}$$

This noise is also low-pass filtered at f_0 until $f_1 = (R_f/R_1)f_0$, where the noise gain reaches unity and it remains until $f_t = 5$ MHz (i.e, the unity-gain frequency of the opamp). Thus, breaking this noise into two portions, and using (9.40) to calculate the first portion, we have

$$\begin{aligned} V_{no2(rms)}^2 &= (265 \times 10^{-9})^2\left(\frac{1}{4R_fC_f}\right) + (24.1 \times 10^{-9})^2\left(\frac{\pi}{2}\right)(f_t - f_1) \\ &= (74.6\ \mu\text{V})^2 \end{aligned} \tag{9.79}$$

Thus, the total output noise is estimated to be

$$V_{no(rms)} = \sqrt{V_{no1(rms)}^2 + V_{no2(rms)}^2} = 77\ \mu\text{V rms} \tag{9.80}$$

It should be noted here that the major source of noise at low frequencies is due to the opamp's voltage noise, $V_n(f)$.

To obtain the SNR for this circuit with a 100-mV rms input level, one can find the output signal mean-squared value and relate that to the output noise value. Alternatively, one can find the equivalent input noise value

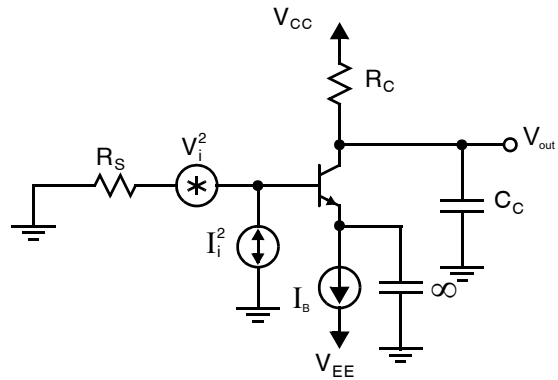


Fig. 9.19 A bipolar common-emitter amplifier.

by dividing the output noise value by the circuit's gain and then relate the input signal and noise mean-squared values. Taking the first approach results in an output signal level of 1 V rms, which gives an SNR of

$$\text{SNR} = 20\log\left(\frac{1\text{ V}}{77\ \mu\text{V}}\right) = 82\text{ dB} \tag{9.81}$$

Note here that using a lower-speed opamp would have reduced the total output noise, as would choosing an opamp with a lower noise voltage. Also note that R_2 contributes to the output noise through both its thermal noise and the noise current of the opamp's positive input. Since its only purpose is to improve the dc offset performance, R_2 should be eliminated in a low-noise circuit (assuming dc offset can be tolerated).

9.4.2 Bipolar Common-Emitter Example

In this example, we consider a bipolar common-emitter amplifier, as shown in Fig. 9.19. Here, we wish to find the optimum bias current to minimize the equivalent input noise of this amplifier (i.e., maximize the signal-to-noise ratio of the amplifier).

We should state that it is assumed here that the collector-current shot noise dominates in the input voltage noise source and the base-current shot noise dominates in the input current noise source. Therefore, the input voltage noise due to the transistor alone is given by

$$V_i^2(f) = 2kT/g_m \tag{9.82}$$

and the input current noise due to the transistor alone is given by

$$I_i^2(f) = 2qI_B \tag{9.83}$$

These assumptions are reasonable since usually 1/f noise is not important for bipolar transistors, and the collector shot noise is usually not a major component of the input current noise source for wideband examples. Also, although the base resistance has temporarily been ignored, it will be taken into account shortly by simply modifying the size of the source resistance.

The first step is to replace the input current noise source by an input voltage noise source. First, note that the gain from the base to the collector (ignoring the transistor output impedance, r_o) is equal to $g_m R_C$, and the impedance looking into the base is r_π. Therefore, the output noise due to $I_i^2(f)$ is given by

$$V_{oi}^2(f) = I_i^2(f)[(R_S \parallel r_\pi)g_m R_C]^2 \tag{9.84}$$

Now the gain from the input voltage noise source to the output is given by

$$A_V = \frac{V_o}{V_i} = \frac{r_\pi}{r_\pi + R_S} g_m R_C \tag{9.85}$$

To represent the output noise due to the input current source by an equivalent input voltage source, we have

$$V_{ieq}^2(f) = \frac{V_{oi}^2(f)}{A_V^2} = \left(\frac{\frac{r_\pi R_S}{(r_\pi + R_S)} g_m R_C}{\frac{r_\pi}{(r_\pi + R_S)} g_m R_C} \right)^2 I_i^2(f) = I_i^2(f) R_S^2 \tag{9.86}$$

We can now replace all noise sources by a single voltage noise source, which includes the noise due to the input voltage noise source and the input current noise source of the transistor, as well as the noise of the source resistor. We have

$$V_{i=\text{total}}^2(f) = V_i^2(f) + V_{ieq}^2(f) + 4kTR_S \tag{9.87}$$

The first two terms represent the noise of the transistor and the third term represents the noise of the source resistance. Using (9.82), (9.83), and (9.86), we have

$$V_{i=\text{total}}^2(f) = \frac{2kT}{g_m} + 2qI_B R_S^2 + 4kTR_S \tag{9.88}$$

Substituting $g_m = qI_C/(kT)$ and $I_B = I_C/\beta$, we find

$$V_{i=\text{total}}^2(f) = \frac{2(kT)^2}{qI_C} + \frac{2qI_C R_S^2}{\beta} + 4kTR_S \tag{9.89}$$

The noise of the transistor base resistor can now be included by replacing R_S with $R_S + r_b$ to obtain

$$V_{i=\text{total}}^2(f) = \frac{2(kT)^2}{qI_C} + \frac{2qI_C(R_S + r_b)^2}{\beta} + 4kT(R_S + r_b) \tag{9.90}$$

Alternatively, (9.90) can be expressed in terms of $g_m = qI_C/(kT)$ as in [Buchwald, 1995]

$$V_{i=\text{total}}^2(f) = 4kT\left[R_s + r_b + \frac{1}{2g_m} + \frac{g_m(R_s + r_b)^2}{2\beta}\right] \tag{9.91}$$

These two equations are good approximations for most applications of bipolar common-emitter amplifiers. Notice in (9.90), the first term models the transistor base-current shot noise and decreases with increased bias current. However, the second term (which models the transistor collector shot noise) increases with increasing collector bias current. Finally, the terms modelling the noise of the source resistor and the transistor base resistor are independent of the transistor bias current. To find the optimum bias current, we differentiate (9.90) with respect to I_C and set the result to zero. This yields the rather simple result of

$$I_{C,\text{opt}} = \frac{kT}{q} \frac{\sqrt{\beta}}{R_S + r_b} \tag{9.92}$$

or, equivalently,

$$g_{m,\text{opt}} = \frac{\sqrt{\beta}}{R_S + r_b} \tag{9.93}$$

EXAMPLE 9.11

Consider the common-emitter amplifier shown in Fig. 9.19, where $R_C = 10\ k\Omega$, $C_C = 1\ pF$, $R_S = 500\ \Omega$, $\beta = 100$, and the base resistance is given by $r_b = 300\ \Omega$. At room temperature, (i.e., 300 °K), find the optimum bias current and the total equivalent input noise.

Solution

Using (9.92), we find

$$I_{C=opt} = 0.026 \times \frac{\sqrt{100}}{500 + 300} = 0.325\ mA \tag{9.94}$$

implying that

$$g_m = \frac{I_C}{V_T} = 12.5\ mA/V \tag{9.95}$$

To find the output noise spectral density, we now use (9.91) to find

$$V^2_{i=total}(f) = 4kT(500 + 300 + 40 + 40) = 1.46 \times 10^{-17}\ V^2/Hz \tag{9.96}$$

Notice that the noise due to the source resistance dominates, even though the source resistance is moderately small. Even for no source resistance, the thermal noise due to the base resistance would still dominate. Thus, for a very low source resistance, the major way to improve the noise is to minimize the base resistance by using larger transistors (or to combine a number of parallel transistors).

Assuming the RC time constant at the collector dominates the frequency response, the noise bandwidth of the amplifier is given by

$$f_x = \frac{1}{4R_CC_C} = 25\ MHz \tag{9.97}$$

resulting in the total input-referred voltage noise given by

$$V_{ni(rms)} = [f_x V^2_{i=total}(f)]^{1/2} = 19.1\ \mu V \tag{9.98}$$

9.4.3 CMOS Differential Pair Example

In this example, we look at the input circuitry of a traditional two-stage CMOS opamp, as shown in Fig. 9.20. As illustrated in Example 9.9, as long as the input stage of a two-stage opamp has some significant gain, it will dominate the overall noise performance of the opamp. Note that each of the transistors have been modelled using an equivalent voltage noise source, as presented in Fig. 9.11. Voltage noise sources are used here since we will be addressing the low-frequency noise performance of this stage. It should be mentioned here that in the following derivations, we have assumed matching between transistor pair Q_1 and Q_2 as well as in pair Q_3 and Q_4.

We start by finding the gains from each noise source to the output node, V_{no}. The gains from V_{n1} and V_{n2} are the same as the gains from the input signals, resulting in

$$\left|\frac{V_{no}}{V_{n1}}\right| = \left|\frac{V_{no}}{V_{n2}}\right| = g_{m1}R_o \tag{9.99}$$

where R_o is the output impedance seen at V_{no}.

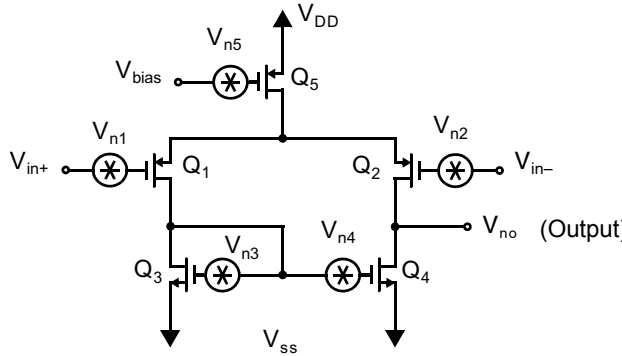


Fig. 9.20 A CMOS input stage for a traditional opamp with MOSFET noise sources shown.

Next, for V_{n3}, notice that the current through this voltage source must be zero because one side is connected to the gate of Q_3 only. Therefore, assuming all other sources are zero, the drain current is unaffected by V_{n3}. This implies that the gate voltage of Q_3 is also unaffected by V_{n3}. Therefore, in the small-signal model, V_{n3} is equal to v_{gs4}, and the gain from V_{n3} to the output is the same as the gain from V_{n4} to the output. Thus, we have

$$\left|\frac{V_{no}}{V_{n3}}\right| = \left|\frac{V_{no}}{V_{n4}}\right| = g_{m3}R_o \tag{9.100}$$

Finally, the noise gain from V_{n5} to the output can be found by noting that it modulates the bias current and the fact that, due to the symmetry in the circuit, the drain of Q_2 will track that of Q_1. As a result, the last gain factor is given by

$$\left|\frac{V_{no}}{V_{n5}}\right| = \frac{g_{m5}}{2g_{m3}} \tag{9.101}$$

Since this last gain factor is relatively small compared to the others, it will be ignored from here on.

Using the gain factors just shown, the output noise value is seen to be given by

$$V_{no}^2(f) = 2(g_{m1}R_o)^2V_{n1}^2(f) + 2(g_{m3}R_o)^2V_{n3}^2(f) \tag{9.102}$$

This output noise value can be related back to an equivalent input noise value, $V_{neq}(f)$, by dividing it by the gain, $g_{m1}R_o$, which results in

$$V_{neq}^2(f) = 2V_{n1}^2(f) + 2V_{n3}^2(f)\left(\frac{g_{m3}}{g_{m1}}\right)^2 \tag{9.103}$$

Thus, for the white noise portion of $V_{n1}(f)$ and $V_{n3}(f)$, we make the substitution

$$V_{ni}^2(f) = 4kT\gamma\left(\frac{1}{g_{mi}}\right) \tag{9.104}$$

resulting in

$$V_{neq}^2(f) = 2 \cdot 4kT\gamma\left(\frac{1}{g_{m1}}\right) + 2 \cdot 4kT\gamma\left(\frac{g_{m3}}{g_{m1}}\right)^2\left(\frac{1}{g_{m3}}\right) \tag{9.105}$$

Key Point: *To minimize thermal noise in a CMOS differential amplifier, the input pair transconductance should be maximized. For a fixed drain current, that implies operation in weak inversion or even subthreshold.*

We see here that the noise contribution of both pairs of transistors is inversely proportional to the transconductance of g_{m1}. *In other wor ds,* g_{m1} *should be made as large as possible to minimize thermal noise* . Also, g_{m3} should be made small to reduce its thermal noise contribution. For a fixed bias current, I_{D5}, this suggests making $V_{eff,1}$ small and $V_{eff,3}$ large. Increasing $V_{eff,3}$ also reduces the output signal swing available, which may adversely effect the maximum SNR achievable in this circuit. However, the signal swing at the output of the first stage in a two-stage opamp is usually relatively small anyway.

Next, we consider the effects of 1/f, or flicker, noise, which normally greatly dominates at low frequencies. First, make the following substitution into (9.103),

$$g_{mi} = \sqrt{2\mu_i C_{ox}\left(\frac{W}{L}\right)_i I_{Di}} \tag{9.106}$$

resulting in

$$V_{ni}^2(f) = 2V_{n1}^2(f) + 2V_{n3}^2(f)\left[\frac{(W/L)_3\mu_n}{(W/L)_1\mu_p}\right] \tag{9.107}$$

Now, letting each of the noise sources have a spectral density given by

$$V_{ni}^2(f) = \frac{K_i}{W_i L_i C_{ox} f} \tag{9.108}$$

we have [Bertails, 1979]

$$V_{ni}^2(f) = \frac{2}{C_{ox} f}\left[\frac{K_1}{W_1 L_1} + \left(\frac{\mu_n}{\mu_p}\right)\left(\frac{K_3 L_1}{W_1 L_3^2}\right)\right] \tag{9.109}$$

Recall that the first term in (9.109) is due to the p-channel input transistors, Q_1 and Q_2, and the second term is due to the n-channel loads, Q_3 and Q_4. We note some points for 1/f noise here:

1. For $L_1 = L_3$, the noise of the n-channel loads dominate since $\mu_n > \mu_p$ and typically n-channel transistors have larger 1/f noise than p-channel transistors (i.e., $K_3 > K_1$).
2. Taking L_3 longer greatly helps due to the inverse squared relationship in the second term of (9.109). This limits the signal swings somewhat, but it may be a reasonable trade-off where low noise is important.
3. The input noise is independent of W_3, and therefore we can make it large if large signal swing is desired at the output (although this will increase the thermal noise of Q_3 and Q_4).
4. Taking W_1 wider also helps to minimize 1/f noise. (Doing so reduces thermal noise, as well.)
5. Taking L_1 longer increases the noise because the second term in (9.109) is dominant. Specifically, this decreases the input-referred noise of the p-channel input transistors, which are not the dominant noise sources, while increasing the input-referred noise of the n-channel load transistors, which are the dominant noise sources.

Finally, we can integrate (9.109) from f_1 to f_2 to find the equivalent input noise value given by

$$V_{ni(rms)}^2 = \frac{2}{C_{ox}}\left[\frac{K_1}{W_1 L_1} + \left(\frac{\mu_n}{\mu_p}\right)\left(\frac{K_3 L_1}{W_1 L_3^2}\right)\right]\ln\left(\frac{f_2}{f_1}\right) \tag{9.110}$$

9.4.4 Fiber-Optic Transimpedance Amplifier Example

The most popular means of detecting light from a fiber-optic cable is to use a transimpedance amplifier, as shown in Fig. 9.21. The light from the fiber cable hits a photodetector, such as a reverse-biased diode. The light produces electron hole carriers in the depletion region of the reverse-biased diode, causing current to flow through the resistor, and therefore the output of the amplifier becomes positive. A popular choice for the first stage of the amplifier is to use a common-source amplifier with a resistor load. In low-noise applications, an active load would be too noisy. Assuming a CMOS transistor is used,[7] the preamp can be modelled as shown in Fig. 9.22. The photodetector is modelled as an input current source along with a parasitic capacitance, C_{in}. Also shown in Fig. 9.22 are the two major noise sources, namely the thermal current noise at the drain of Q_1 and the thermal noise from the feedback resistor, R_F. The 1/f noise of the transistor is ignored because it is assumed that the circuit is high speed, such that thermal noise dominates. The second (and perhaps subsequent) stage is modelled by an amplifier that has a gain of A_2. The noise due to this second stage is also ignored since the noise sources in the second stage are not amplified by the gain of the first stage.

A simplified small-signal model of this preamplifier, used for noise analysis, is shown in Fig. 9.23. The only parasitic capacitance considered in the transistor model is C_{gs}, since the gain of the first stage is only moderate,

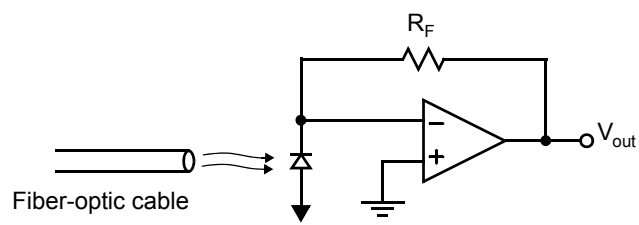


Fig. 9.21 A fiber-optic transresistance preamp.

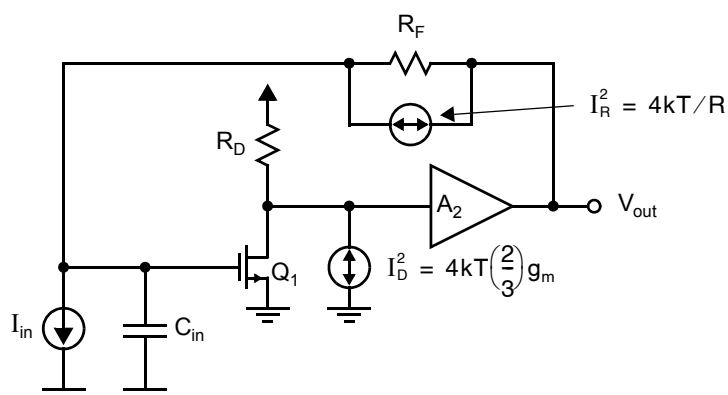


Fig. 9.22 A simplified model for a CMOS fiber-optic preamp.

7. Fiber-optic preamps have also been realized using JFET transistors as well, but since their thermal noise model is identical to the noise model of the CMOS transistor, the analysis would be almost unchanged.

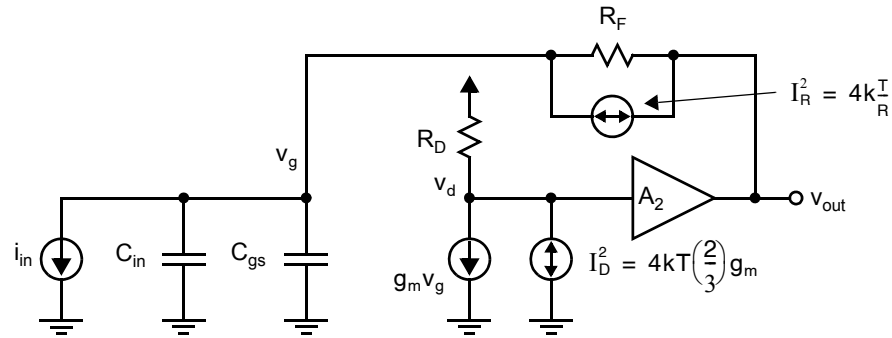


Fig. 9.23 A simplified small-signal model used for noise analysis.

due to the resistor load, and therefore we can assume that C_{gd} and C_{db} can be ignored. The transfer function from i_{in} to the output is found by using nodal analysis to be

$$\frac{V_{out}(s)}{I_{in}(s)} = R_F\left(\frac{A_V}{1 + A_V}\right)\frac{1}{1 + s\left(\frac{R_F C_T}{1 + A_V}\right)} \tag{9.111}$$

where $A_V = A_1 A_2$ is the total voltage gain of the preamp and $C_T = C_{gs} + C_{in}$. It is also found that this is the same transfer function from I_R^2 to the output. Thus, the noise current source I_R^2 can be replaced by an input current noise source having a spectral density of $4kT/R_F$. In a typical design, one would choose R_F as large as possible to limit this noise. However, this choice is constrained by the bandwidth requirements of passing the signal frequencies. From (9.111), we have the –3-dB frequency of the amplifier, given by

$$\omega_{-3\text{ dB}} = \frac{1 + A_V}{R_F C_T} \tag{9.112}$$

For a given amplifier gain and detector capacitance, R_F is chosen using (9.112) to place the –3-dB frequency as small as possible without substantially attenuating the signals. This makes the dominant node for determining stability the input node. The bandwidth of the second amplifier must be substantially greater than that of the input node to guarantee stability. Unfortunately, this constraint greatly amplifies the thermal noise due to input transistor Q_1, as we will see next.

The gain from the noise source I_D^2 to the output is found by using nodal analysis to be

$$\frac{V_{out}(s)}{I_D(s)} = \frac{1}{g_m}\frac{A_V}{1 + A_V}\frac{1 + sR_F C_T}{1 + s\left[\frac{R_F C_T}{(1 + A_V)}\right]} \tag{9.113}$$

At low frequencies, this gain is approximately given by $1/g_m$, whereas at high frequencies it is as much as $1 + A_V$ times greater. Thus, the high-frequency output noise due to Q_1 is much greater than the low-frequency noise due to Q_1. Furthermore, the only bandwidth limitation of this noise at high frequencies is due to the finite bandwidth of the second amplifier, A_2. From the discussion on compensation in Chapter 5, we know the second pole frequency of the amplifier must be almost four times greater than the closed-loop –3-dB frequency when lead compensation is not used.[8]

8. Lead compensation could possibly be achieved by placing a capacitor in parallel with R_F.

Continuing, in order to refer the noise due to Q_1 back to the input, we use (9.111) and (9.113) to obtain

$$I_{iD}^2(f) = I_D^2(f)\left|\frac{V_{out}(j\omega)}{I_D(jw)}\right|^2\left|\frac{I_{in}(j\omega)}{V_{out}(jw)}\right|^2 = I_D^2(f)\frac{1+\omega^2(R_FC_T)^2}{(g_mR_F)^2} \tag{9.114}$$

We use

$$I_D^2(f) = 4kT\gamma g_m \tag{9.115}$$

and we add to this the input noise source that models the noise of the feedback resistor to obtain the total input-referred noise, given by

$$I_i^2(f) = \frac{4kT}{R_F} + 4kT\gamma\frac{1+\omega^2(R_FC_T)^2}{g_mR_F^2} \tag{9.116}$$

Notice that the second term starts to quickly increase at a frequency of

$$\omega_z = \frac{1}{R_FC_T} = \frac{\omega_{-3\,dB}}{1+A_V} \tag{9.117}$$

which is a relatively low frequency.

Continuing, normally $\gamma/(g_mR_F) << 1$, and we have

$$\begin{aligned} I_i^2(f) &= \frac{4kT}{R_F}\left(1+\frac{\gamma}{g_mR_F}\right) + 4kT\gamma\frac{\omega^2C_T^2}{g_m} \\ &\cong \frac{4kT}{R_F} + 4kT\gamma\frac{\omega^2C_T^2}{g_m} \end{aligned} \tag{9.118}$$

Using the facts that

$$g_m = \mu_nC_{ox}\frac{W}{L}V_{eff} \tag{9.119}$$

and that

$$C_{gs} = \frac{2}{3}C_{ox}WL \tag{9.120}$$

we can write (9.118) as

$$I_i^2(f) \cong \frac{4kT}{R_F} + \frac{16}{9}kT\frac{L^2}{\mu_nV_{eff}}\omega^2\frac{(C_{gs}+C_{in})^2}{C_{gs}} \tag{9.121}$$

Normally, V_{eff} is taken as large as possible given power-supply-voltage and power-dissipation constraints. The only parameter left for the designer to choose is the width of the input transistor and, therefore, C_{gs}. It can be shown (by differentiating the second term of (9.121) with respect to C_{gs} and then setting the result to zero) that the second term of (9.121) is minimized by the choice $C_{gs} = C_{in}$, in which case we have

$$I_i^2(f) \cong 4kT\left(\frac{1}{R_F} + \frac{4}{9}\frac{L^2}{\mu_nV_{eff}}\omega^2C_{in}\right) \tag{9.122}$$

9.5 DYNAMIC RANGE PERFORMANCE

Analog circuits generally have many different performance criteria that must be met so that overall system performance meets specifications. Whereas noise limits the value of the smallest useful signals, linearity limits the value of the largest useful signals that can occur in the circuit. Thus, we find that linearity and noise together determine

the useful dynamic range of a circuit. It should be noted here that the dynamic range of integrated analog circuits is often a crucial measure since its value is often low and can seriously impair system performance.

9.5.1 Total Harmonic Distortion (THD)

If a sinusoidal waveform is applied to a linear time-invariant system, it is well known that the output will also be a sinusoidal waveform at the same frequency, but possibly with different magnitude and phase values. However, if the same input is applied to a nonlinear system, the output signal will have frequency components at harmonics of the input waveform, including the fundamental (or first) harmonic. For example, if a 1-MHz sinusoidal input signal is used, the output signal will have power at the fundamental, 1 MHz, as well as at the harmonic frequencies, 2 MHz, 3 MHz, 4 MHz, and so on.

Specifically, consider a nonlinear system[9] with an input signal, $v_{in}(t)$, and an output signal, $v_o(t)$. The output signal can be written as a Taylor series expansion of the input signal:

$$v_o(t) = a_1 v_{in}(t) + a_2 v_{in}^2(t) + a_3 v_{in}^3(t) + a_4 v_{in}^4(t) + \cdots \tag{9.123}$$

Here, the linear term is a_1, whereas a_2, a_3, and a_4 characterize the second-, third-, and fourth-order distortion terms, respectively. In fully differential circuits, all even terms (i.e., a_2 and a_4) are small, so typically a_3 dominates and we approximate $v_o(t)$ as

$$v_o(t) \cong a_1 v_{in}(t) + a_3 v_{in}^3(t) \tag{9.124}$$

If $v_{in}(t)$ is a sinusoidal signal given by

$$v_{in}(t) = A\cos(\omega t) \tag{9.125}$$

the output signal can be shown to be approximated by

$$v_o(t) \cong a_1 A\cos(\omega t) + \frac{a_3}{4}A^3[3\cos(\omega t) + \cos(3\omega t)] \tag{9.126}$$

where we see a fundamental term and a third-harmonic term. Defining H_{D1} and H_{D3} to be the amplitudes of the fundamental and third-harmonic terms, respectively, we can write

$$v_o(t) \equiv H_{D1}\cos(\omega t) + H_{D3}\cos(3\omega t) \tag{9.127}$$

Since, typically, $(3/4)a_3A^3 \ll a_1A$, one usually approximates the linear component of the output signal as

$$H_{D1} = a_1 A \tag{9.128}$$

and the third-harmonic term as

$$H_{D3} = \frac{a_3}{4}A^3 \tag{9.129}$$

Finally, we see that the third-order distortion term results in power at the third harmonic frequency and the ratio of H_{D3}/H_{D1} is defined as the third-order harmonic distortion ratio, given by

$$HD_3 \equiv \frac{H_{D3}}{H_{D1}} = \left(\frac{a_3}{a_1}\right)\left(\frac{A^2}{4}\right) \tag{9.130}$$

The *total harmonic distortion* (THD) of a signal is defined to be the ratio of the total power of all second and higher harmonic components to the power of the fundamental for that signal. In units of dB, THD is found using

9. We assume here that the nonlinear system is memoryless and time invariant. Unfortunately, circuits are not memoryless, and a Volterra series should be used; however, this assumption simplifies the analysis and usually results in good approximations for distortion figures.

the following relation:

$$\text{THD} = 10\log\left(\frac{H_{D2}^2 + H_{D3}^2 + H_{D4}^2 + \cdots}{H_{D1}^2}\right) \tag{9.131}$$

Sometimes THD is presented as a percentage value. In this case,

$$\text{THD} = \frac{\sqrt{H_{D2}^2 + H_{D3}^2 + H_{D4}^2 + \cdots}}{H_{D1}} \times 100\,\% \tag{9.132}$$

For example, a 0.1-percent THD value implies that the amplitude of the fundamental is 1,000 times larger than the amplitude of the harmonic components. This 0.1-percent THD value is equivalent to a value of –60 dB THD. It should also be noted that the THD value is almost always a function of the amplitude of the input (or output) signal level, and thus the corresponding signal amplitude must also be reported. Also, for practical reasons, typically the power of only the first few harmonics (say, the first 5) are included since the distortion components usually fall off quickly for higher harmonics.

The total harmonic distortion of a circuit deteriorates as the applied signal amplitudes are increased. This is intuitive since we regularly employ linear circuit models for the analysis of small signals. For example, in the case of a fully-differential circuit where the dominant distortion is H_{D3}, the total harmonic distortion is approximately equal to HD_3, which in (9.130) is shown to be proportional the input amplitude squared, A^2.

One difficulty with the use of THD in reporting circuit performance is that often the harmonic components fall outside of the circuit's usable bandwidth, and thus the THD value is falsely improved. For example, if a 21-MHz low-pass filter is being tested, then a 5-MHz input signal results in the 10-MHz, 15-MHz, and 20-MHz components falling within the passband, whereas higher distortion components will be attenuated by the filter. However, circuit linearity is often worse when higher input signal frequencies are applied due to nonlinear capacitances or, worse, nonlinear signal cancellation. Thus, it is useful to measure circuit linearity with input signals near the upper edge of the passband. Unfortunately, if a 20-MHz input signal is applied to the example low-pass filter, the second and higher harmonic components lie in the stopband and the THD value will indicate much better linearity than would occur for a practical application. Moreover, the testing of a narrowband filter always has this THD measurement difficulty since the use of an input frequency in the passband will result in harmonics that will certainly be attenuated by the filter's stopband. In summary, a THD measurement is straightforward to perform but does not work well in the important test of high-frequency signals near the upper passband limit of the circuit. Fortunately, as we will see next, one can use intermodulation tests to measure circuit linearity near the upper passband edge.

> ***Key Point:*** *Total harmonic distortion (THD) measures the ratio of fundamental signal power to that of all harmonics. A THD measurement is straightforward to perform only when the fundamental signal frequency is well below the circuit's upper passband limit.*

EXAMPLE 9.12

Calculate the THD value of a signal in which the fundamental component is at 1 MHz and has a power level of –10 dBm, and the only significant harmonics are the first three above the fundamental. These harmonics, at 2 MHz, 3 MHz, and 4 MHz, are measured to have power levels of –53 dBm, –50 dBm, and –56 dBm, respectively.

Solution

Recalling that 0 dBm refers to a 1-mW signal, then the power of the fundamental component is 100 μW, whereas the power of the first three harmonics are 5 nW, 10 nW, and 2.5 nW, respectively. Thus, the total power of the harmonics is equal to

$$P_{harm} = 5 + 10 + 2.5 = 17.5 \text{ nW} \tag{9.133}$$

As a result, the THD is calculated to be

$$\text{THD} = 10\log\left(\frac{17.5\ \text{nW}}{100\ \mu\text{W}}\right) = -37.6\ \text{dB} \tag{9.134}$$

9.5.2 Third-Order Intercept Point (IP3)

Here, we introduce the concept of the *third-order intercept point* (IP3) as a measure for the third-order distortion component [Carson, 1990]. Unfortunately, as just noted, distortion terms for a single sinusoidal input often lie outside of a circuit's bandwidth, and thus we resort to an intermodulation test to move the distortion term back near the frequency of the input signals. We focus here on third-order distortion since, for fully differential circuits, the even-order distortion components are ideally zero and, thus, third-order distortion typically dominates. However, in cases where second-order distortion dominates, a similar concept is also possible and is referred to as the *second-order intercept point* (IP2).

Consider an intermodulation test, where the input signal consists of two equally sized sinusoidal signals and is written as

$$v_{in}(t) = A\cos(\omega_1 t) + A\cos(\omega_2 t) \tag{9.135}$$

Presuming the input–output relationship of (9.123), the output signal can be shown to be approximated by

$$\begin{aligned} v_o(t) \cong & \left(a_1 A + \frac{9a_3}{4}A^3\right)[\cos(\omega_1 t) + \cos(\omega_2 t)] \\ & + \frac{a_3}{4}A^3[\cos(3\omega_1 t) + \cos(3\omega_2 t)] \\ & + \frac{3a_3}{4}A^3[\cos(2\omega_1 t + \omega_2 t) + \cos(2\omega_2 t + \omega_1 t)] \\ & + \frac{3a_3}{4}A^3[\cos(\omega_1 t - \Delta\omega t) + \cos(\omega_2 t + \Delta\omega t)] \end{aligned} \tag{9.136}$$

where $\Delta\omega$ is defined to be the difference between the input frequencies (i.e., $\Delta\omega \equiv \omega_2 - \omega$) which we assume to be small. Here, we see that the first line of (9.136) is the fundamental components, the second line shows the levels at three times the fundamentals, the third line also describes distortion at nearly three times the fundamentals, and the fourth line describes the distortion levels at two new frequencies that are close to the input frequencies (slightly below ω_1 and slightly above ω_2). As a result, for a narrowband or low-pass circuit, these two new distortion components (due to third-order distortion) fall in the passband and can be used to predict the third-order distortion term.

Using the same notation as in the harmonic distortion case, we have the intermodulation distortion levels given by

$$I_{D1} = a_1 A \tag{9.137}$$

and

$$I_{D3} = \frac{3a_3}{4}A^3 \tag{9.138}$$

The ratio of these two is the third-order intermodulation value, given by

$$ID_3 = \frac{I_{D3}}{I_{D1}} = \left(\frac{a_3}{a_1}\right)\left(\frac{3A^2}{4}\right) \tag{9.139}$$

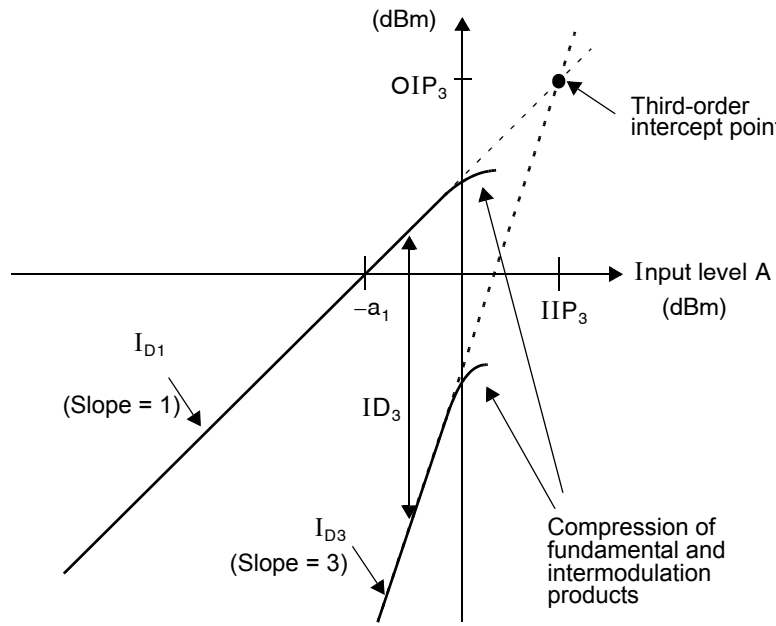


Fig. 9.24 Graphical illustration of the third-order intercept-point. IIP_3 and OIP_3 are the input and output third-order intercept points, respectively. They cannot be measured directly due to compression of the fundamental and intermodulation products at high-power levels.

From (9.137) and (9.138), note that, as the amplitude of the input signal, A, is increased, the level of the fundamental rises linearly, whereas the I_{D3} rises in a cubic fashion. For example, if A is increased by 1 dB, then I_{D1} also increases by 1 dB while I_{D3} increases by 3 dB. Thus, when the fundamental and intermodulation levels are plotted in a dB scale against the signal amplitude, A, the two curves are linear, but with different slopes, as shown in Fig. 9.24. The third-order intercept point is defined to be the intersection of these two lines. Note, however, that as the signal amplitude rises, the linear relationships of I_{D1} and I_{D3} are no longer obeyed due to the large amount of nonlinearities violating some of our original assumptions. For example, in (9.136), we ignored the cubic A term in estimating the fundamental level, I_{D1}. Also, other distortion terms that were ignored now become important. As a result, it is impossible to directly measure the third-order intercept point, and thus it must be extrapolated from the measured data. The third-order intercept point results in two values—IIP_3 and OIP_3, which are related to the input and output levels, respectively. If the linear gain term, a_1, is unity (or, equivalently, 0 dB), then $\mathrm{IIP}_3 = \mathrm{OIP}_3$. However, if a_1 is not unity, one must be careful to state which of the two intercept points is being reported.

Knowledge of the third-order intercept point is quite useful in determining what signal level should be chosen to achieve a desired intermodulation ratio, ID_3. Specifically, we see from (9.139) that ID_3 improves by 2 dB for every 1 dB of signal level decrease (since it is related to A^2) and that OIP_3 is defined to be the I_{D1} point where $\mathrm{ID}_3 = 0$ dB. Thus, we have the simple relationship (all in decibels)

$$\mathrm{OIP}_3 = I_{D1} - \frac{\mathrm{ID}_3}{2} \tag{9.140}$$

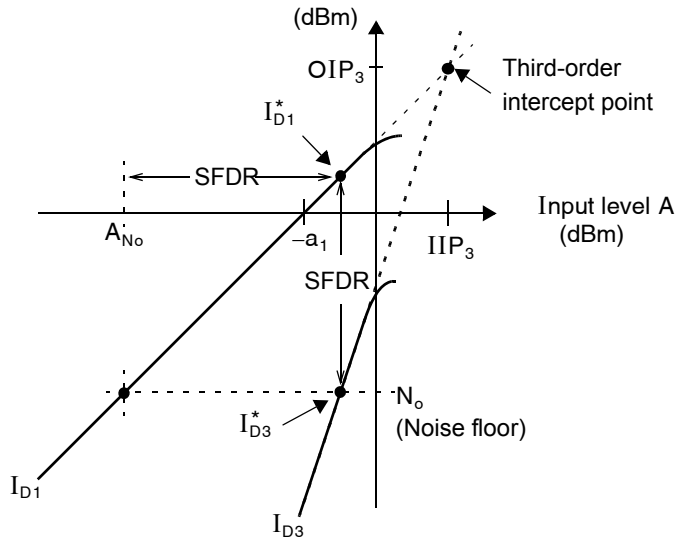


Fig. 9.25 Graphical illustration of spurious-free dynamic range (SFDR).

EXAMPLE 9.13

If $\mathrm{OIP}_3 = 20$ dBm, what output-signal level should be used such that the third-order intermodulation products are 60 dB below the fundamental?

Solution

Using (9.140) with $\mathrm{ID}_3 = -60$ dB, we have

$$\mathrm{I}_{\mathrm{D1}} = \mathrm{OIP}_3 + \frac{\mathrm{ID}_3}{2} = -10 \text{ dBm} \tag{9.141}$$

Thus, an output level of -10 dBm should be used.

9.5.3 Spurious-Free Dynamic Range (SFDR)

Spurious-free dynamic range (SFDR) is defined to be the signal-to-noise ratio when the power of the distortion equals the noise power. In an intermodulation test, the third order intermodulation products are the dominant distortion. In Fig. 9.25, the circuit's total output noise power is shown along the vertical axis as $\mathrm{N_o}$. If a low enough signal level is used, $\mathrm{I_{D3}}$ will be well below the noise floor. However, since $\mathrm{I_{D3}}$ rises 3 dB for every 1 dB of signal-level increase, there will soon be a point where $\mathrm{I_{D3}}$ is equal to the noise power. As the figure shows, SFDR is defined to be the output SNR ratio when $\mathrm{I_{D3}}$ is equal to $\mathrm{N_o}$. Alternatively, one can measure SFDR using the input-signal levels as the difference between the level that results in $\mathrm{I_{D3}} = \mathrm{N_o}$ and the level $\mathrm{A_{No}}$ that results in a fundamental output level equal to $\mathrm{N_o}$.

To find a formula relating SFDR, OIP_3, and $\mathrm{N_o}$ (all in dB units), we first note that

$$\mathrm{SFDR} = \mathrm{I}^*_{\mathrm{D1}} - \mathrm{N_o} = \mathrm{I}^*_{\mathrm{D1}} - \mathrm{I}^*_{\mathrm{D3}} \tag{9.142}$$

where $\mathrm{I}^*_{\mathrm{D1}}$ and $\mathrm{I}^*_{\mathrm{D3}}$ refer to the output and distortion levels when $\mathrm{I_{D3}} = \mathrm{N_o}$, as shown in Fig. 9.25. Since the units are assumed to be in dB, we also have, from (9.139),

$$\mathrm{ID}_3 = \mathrm{I_{D3}} - \mathrm{I_{D1}} \tag{9.143}$$

Using (9.140), we have

$$\mathrm{OIP}_3 = \mathrm{I}^*_{\mathrm{D1}} - \frac{(\mathrm{N_o} - \mathrm{I}^*_{\mathrm{D1}})}{2} \tag{9.144}$$

and substituting in $\mathrm{I}^*_{\mathrm{D1}} = \mathrm{SFDR} + \mathrm{N_o}$ from (9.142) and rearranging, we have

$$\mathrm{SFDR} = \frac{2}{3}(\mathrm{OIP}_3 - \mathrm{N_o}) \tag{9.145}$$

EXAMPLE 9.14

At an input-signal level of 0 dBm, an intermodulation ratio of –40 dB was measured in a circuit with a gain of 2 dB. Calculate the value of the input and output third-order intercept points. What input-signal level should be applied if one wants an intermodulation ratio of –45 dB? If the noise power at the output is measured to be –50 dBm, what is the expected SFDR, and what output-signal level does it correspond to?

Solution

With an input level of 0 dBm and a gain of 2 dB, the output level is $\mathrm{I_{D1}} = 2$ dBm with a measured value of $\mathrm{ID}_3 = -40$ dB. Using (9.140), we have

$$\mathrm{OIP}_3 = \mathrm{I_{D1}} - \frac{\mathrm{ID}_3}{2} = 22 \text{ dBm} \tag{9.146}$$

whereas the IIP_3 is 2 dB lower (i.e., $\mathrm{IIP}_3 = 20$ dBm).

For signal levels corresponding to an intermodulation of $\mathrm{ID}_3 = 45$ dB, we have

$$\mathrm{I_{D1}} = \mathrm{OIP}_3 + \frac{\mathrm{ID}_3}{2} = 22 - \frac{45}{2} = -0.5 \text{ dBm} \tag{9.147}$$

However, this value is the level of the output signal, so the input signal should be 2 dB lower or, equivalently, the input-signal level should be at –2.5 dBm. Note that this result could have been obtained more directly by noting that a 5-dB improvement in distortion was desired, implying that the signal level should be decreased by (5 dB)/2.

Finally, we use (9.145) to find

$$\mathrm{SFDR} = \frac{2}{3}(22 + 50) = 48 \text{ dB} \tag{9.148}$$

from which the output level is given by

$$\mathrm{I}^*_{\mathrm{D1}} = \mathrm{SFDR} + \mathrm{N_o} = -2 \text{ dBm} \tag{9.149}$$

Thus, when the output level is at -2 dBm, the third-order intermodulation power equals the noise power. If the output level is increased, the distortion products will increase and limit the dynamic-range performance. However, if the output level is decreased, the distortion products will be buried below the noise floor, and the noise will limit dynamic-range performance. Therefore, optimum dynamic-range performance is obtained at an output level of -2 dBm. However, note that the dynamic-range performance is actually 3 dB below the SFDR value since the dynamic range is based on the ratio of the signal power to the distortion plus the noise power (which are both equal at this point).

9.5.4 Signal-to-Noise and Distortion Ratio (SNDR)

The signal-to-noise and distortion ratio (SNDR) of a signal is defined as the ratio of the signal power to the total power in all noise and distortion components. Usually measured with a single-tone input, the signal power is the amplitude of the fundamental, and the distortion includes all harmonics.

$$\text{SNDR} = 10 \log \left(\frac{V_f^2}{N_o + V_{h2}^2 + V_{h3}^2 + V_{h4}^2 + \cdots} \right) \tag{9.150}$$

Unlike SFDR, SNDR is a function of signal amplitude. For small signal power levels, the harmonics are negligible relative to the noise, and the SNDR is well approximated by the SNR, $\text{SNDR} \cong 10 \log (V_f^2 / N_o)$. In this circumstance, assuming the noise power remains constant as the signal power varies, we see an increasing SNDR with increasing signal power. By contrast, at high signal power levels the harmonic distortion terms dominate the denominator of (9.150), at which point SNDR is approximately equal to $1/\text{THD}$, $\text{SNDR} \cong 10 \log (V_f^2 / V_{h2}^2 + V_{h3}^2 + V_{h4}^2 + \cdots)$. Since harmonic power levels increase faster than the fundamental power, SNDR decreases at high signal powers. Hence, SNDR varies with signal amplitude as shown in Fig. 9.26. SNDR is maximized at a particular input amplitude. The maximal value SNDR_{max} is, like SFDR, a useful metric for quantifying the dynamic range of a circuit.

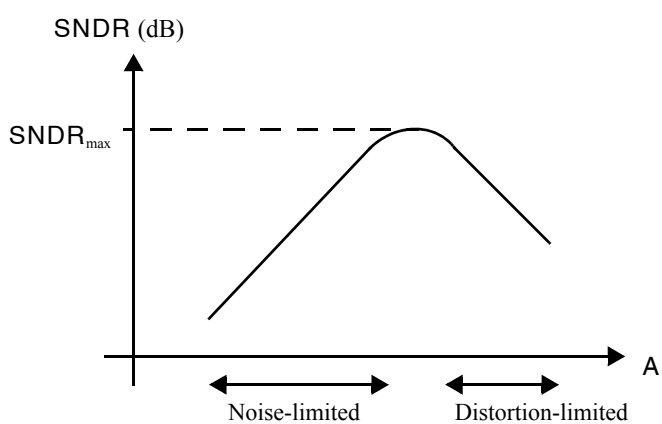


Fig. 9.26 Variation of SNDR with signal amplitude in an analog circuit.

9.6 KEY POINTS

- This chapter focuses on noise inherent to the physical operation of electronic components, not noise from outside interferences. Such noise sources are generally zero-mean. [p. 363]
- Although the instantaneous value of a noise voltage or current cannot be determined by analysis, the rms value obtained by time-averaging its square provides a normalized measure of the noise signal's power. [p. 364]
- SNR is the ratio of signal power to noise power, both normalized to the same impedance and expressed in dB. [p. 365]
- Units of dBm express an average power level on a logarithmic scale normalized to 1 mW, so that 1 mW is 0 dBm and 0.1 mW is 10 dBm. Rms voltages or currents are also sometimes expressed in dBm by assuming that they are applied to some agreed-upon reference load, usually either 50 or 75 Ohms. [p. 365]
- The power of independent (uncorrelated) noise sources superimpose linearly, which means their voltages and currents superimpose in a root-sum-of-squares fashion. As a result, to lower overall noise, one must focus on the largest noise contributors. [p. 366]
- The mean square value of a noise signal is the integral of its spectral density over all frequencies. [p. 369]
- The frequency-domain representation of a noise signal is its spectral density: the average normalized power that may be measured within a 1-Hz bandwidth. [p. 368]
- White noise has a spectral density that is constant over all frequencies. [p. 369]
- Flicker noise, or 1/f-noise, has a spectral density proportional to 1/f. Hence, it falls off at –10 dB/decade. [p. 370]
- Uncorrelated noise signals remain uncorrelated, even when filtered by a circuit's magnitude response. [p. 372]
- Passing a noise signal through a circuit shapes its root spectral density by the circuit's magnitude response. [p. 372]
- The noise bandwidth of a given filter is equal to the frequency span of a brick-wall filter that has the same rms output noise as the given filter has when white noise is applied to both filters, assuming the same peak gain for both filters. [p. 373]
- The noise bandwidth of a first-order lowpass filter is $\pi/2$ times the filter's –3dB bandwidth. [p. 375]
- To determine the frequency region or regions that contribute the dominant noise, lower a 1/f noise line until it touches the spectral density curve. The total noise can be approximated by the noise in the vicinity of the 1/f line. [p. 377]
- Resistors exhibit thermal noise modeled either by a voltage source in series with the resistor having a white noise spectral density $4kTR$, or a Norton equivalent current source having a white noise spectral density $4kT/R$. [p. 378]
- Diodes exhibit shot noise modeled by a current source having a white noise spectral density $2qI_D$. [p. 378]
- At low frequencies, 1/f noise is dominant in MOSFET circuits. It is minimized by using large gate areas and p-channel transistors wherever possible. [p. 380]
- Bipolar transistor noise is typically dominated by the thermal noise of the series base resistance, and shot noise in the base-current junction. [p. 380]
- MOSFET thermal noise is simply that of a resistor with value r_{ds} when in triode. In active mode, thermal noise is modeled by a white current noise source between drain and source with spectral density $8kTg_m/3$, for a long channel device, and somewhat higher for short-channel devices. [p. 381]
- In a circuit comprising only capacitors and resistors (including triode MOSFETs) but no active devices, the squared rms noise on each capacitor is kT/C. [p. 383]
- The input-referred noise of a circuit, if applied to the input of a noiseless copy of the circuit, results in the exact same output noise as when all of the circuit's noise sources were present. It is found by dividing the observed output noise by the gain of the circuit. [p. 384]

- The input-referred noise of a circuit, if applied to the input of a noiseless copy of the circuit, results in the exact same output noise as when all of the circuit's noise sources were present. It is found by dividing the observed output noise by the gain of the circuit. [p. 384]
- Sampling an analog voltage on a capacitor captures both the signal component and noise with a mean-square value of kT/C, assuming a purely passive sampling network. [p. 384]
- The noise performance of a circuit at a particular frequency is usually reported as its input-referred noise spectrum, equal to the circuit's output noise spectrum divided by its magnitude response, or equivalently by its noise figure. [p. 387]
- To minimize thermal noise in a CMOS differential amplifier, the input pair transconductance should be maximized. For a fixed drain current, that implies operation in weak inversion or even subthreshold. [p. 394]

9.7 REFERENCES

J. C. Bertails. "Low-Frequency Noise Considerations for MOS Amplifiers Design," *IEEE J. of Solid-State Circuits,* Vol. SC-14, no. 4, pp. 773–776, August, 1979.

A. Buchwald and K. Martin. *Integrated Fiber-Optic Receivers*. Kluwer Academic Publishers, Norwell, Massachusetts, 1995.

S. Franco. *Design with Operational Amplifiers and Analog Integrated Circuits*. McGraw-Hill, New York, Chapter 7, 2002.

P. R. Gray and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits.* John Wiley & Sons, New York, 1993.

J. B. Johnson. *Phys. Rev.*, Vol. 32, pp. 97–109, 1928.

E. J. Kennedy. *Operational Amplifier Circuits: Theory and Applications.* Holt, Rinehart, & Winston, New York, 1988.

A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, Reading, Massachusetts, 1989.

C. D. Motchenbacher and J. A. Connelly. *Low Noise Electronic System Design*. John Wiley & Sons, New York, 1993.

H. Nyquist. *Phys. Rev.*, Vol. 32, pp. 110–113, 1928.

W. Schottky. *Ann. Phys. (Leipzig)*, Vol. 57, pp. 541–567, 1918.

M. Steyaert, Z. Y. Chang, and W. Sansen. "Low-Noise Monolithic Amplifier Design: Bipolar versus CMOS," *Analog Integrated Circuits and Signal Processing 1*. Kluwer Academic Publishers, Boston, pp. 9–19, September 1991.

A. Van der Ziel. *Noise in Solid State Devices and Circuits*. John Wiley & Sons, New York, 1986.

9.8 PROBLEMS

Unless otherwise stated, assume dBm values are referenced to $50\ \Omega$.

9.8.1 Section 9.1: Time-Domain Analysis

9.1 If a signal is measured to have $V_{(rms)}$ volts, what is the difference in dB if it is expressed in dBm referenced to $50\ \Omega$ as opposed to being referenced to $75\ \Omega$?

9.2 A sinusoidal signal of 150 mV rms is in series with a noise voltage.

a. If the noise voltage is 1 mV rms, what is the signal to noise ratio expressed in dB?

b. If a SNR of 70 dB is desired, how small must the noise voltage be made?

9.3 A noise current has a power of –30 dBm referenced to a $50\ \Omega$ load. What is its rms value in Amperes?

9.4 What signal level is required, in $V_{(rms)}$, to achieve a SNR of 62 dB in the presence of –35 dBm noise?

9.5 Consider the sum of two noise sources of values –20 dBm and –23 dBm. Find the total noise power in dBm for the cases in which the two noise sources are (*a*) uncorrelated, (*b*) $C = 0.3$, (*c*) $C = +1$, and (*d*) $C = -1$.

9.8.2 Section 9.2: Frequency-Domain Analysis

9.6 The output noise of a circuit is measured to be -40 dBm around 100 kHz when a resolution bandwidth of 30 Hz is used. What is the expected dBm measurement if a resolution bandwidth of 10 Hz is used? Find the root spectral density in $V/\sqrt{Hz}$.

9.7 At 0.1 Hz, a low-frequency measurement has a noise value of -60 dBm when a resolution bandwidth of 1 mHz is used. Assuming 1/f noise dominates, what would be the expected noise value (in dBm) over the band from 1 mHz to 1 Hz?

9.8 Using the 1/f tangent principle, estimate the total noise above 0.1 Hz for the spectral density shown in Fig. 9.3.

9.9 Consider a bandpass amplifier that has equivalent input noise root spectral density and amplifier response, as shown in Fig. P9.9. Sketch the root spectral density for the output signal. Estimate the total output rms noise value by applying the 1/f tangent principle.

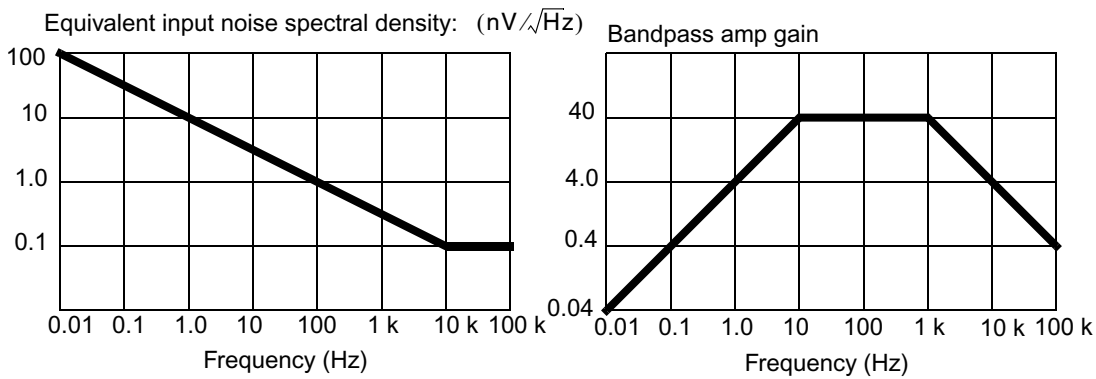


Fig. P9.9

9.10 Consider the noise root spectral density of the signal shown in Fig. P9.10. Find the total rms noise value using a graphical approach for 0.01 to ∞ Hz. Compare your result with that obtained when using the 1/f tangent principle.

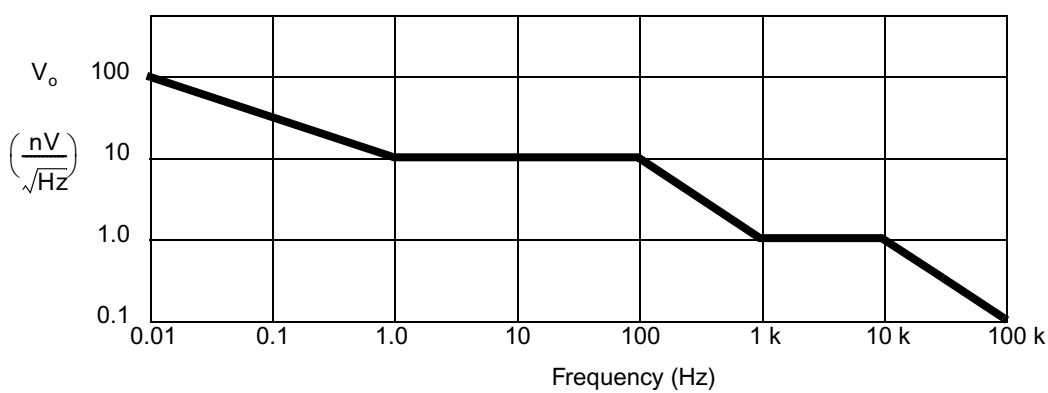


Fig. P9.10

9.11 The plot in Fig. P9.11 is obtained from a spectrum analyzer showing the spectrum of a sinusoidal signal in noise with a resolution bandwidth of 10 kHz.

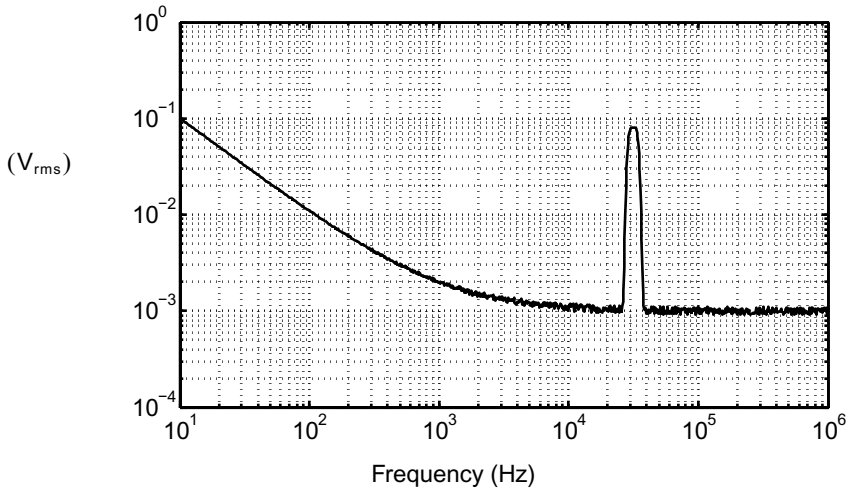


Fig. P9.11

a. Find the spectral density of the white noise component, V_{nw}, in $V/\sqrt{Hz}$.

b. Estimate the coefficient for the 1/f noise component, k_v.

c. Find the 1/f noise corner frequency.

d. What is the total integrated noise power from 10 Hz to 1 MHz expressed in both V_{rms} and dBm referenced to 50 Ω?

e. Find the SNR considering all noise from 10 Hz to 1 MHz.

f. Sketch the spectral plot that would be obtained if the resolution bandwidth was decreased to 1 kHz.

9.12 Repeat Problem 9.11 for the plot in Fig. P9.12.

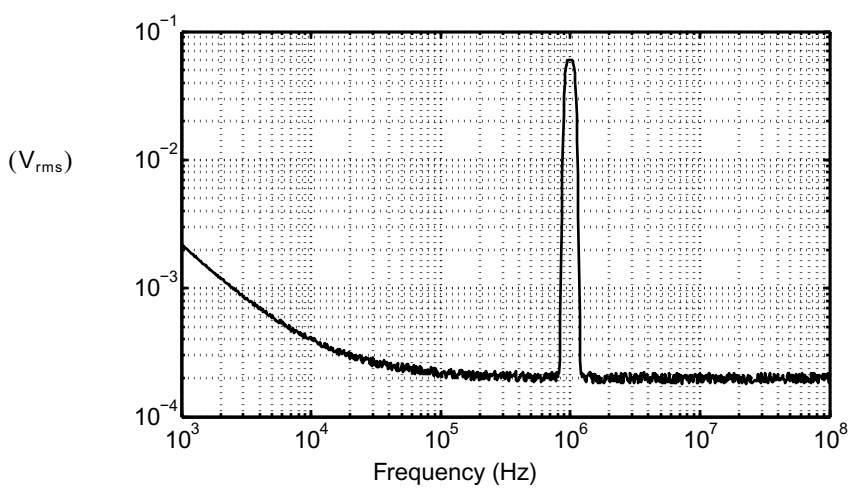


Fig. P9.12

9.13 We saw in equation (9.41) that the noise bandwidth of a first-order, low-pass filter is $(\pi/2)f_0$. Show that the noise bandwidth of a second-order, low-pass filter given by

$$A(s) = \frac{A_o}{\left(1 + \frac{s}{(2\pi f_0)}\right)^2}$$

is equal to $(\pi/4)f_o$.

9.8.3 Section 9.3: Noise Models for Circuit Elements

9.14 Show that, when two resistors of values R_1 and R_2 are in series, their noise model is the same as a single resistance of value $R_1 + R_2$. Repeat the problem for parallel resistances.

9.15 Sketch the spectral density of voltage noise across a 100-pF capacitor when it is in parallel with a $1\text{-k}\Omega$ resistor. Make another sketch for the same capacitor but with a $1\text{-M}\Omega$ resistance in parallel. What can you say about the area under the curves of the two sketches?

9.16 In Fig. P9.16 v_i is a 0.2 V_{rms} sinusoid. Find the SNR at v_o as a function of the frequency of v_i.

9.17 Consider the circuit shown in Fig. P9.17, where the opamp has a unity-gain frequency of 1 MHz and equivalent input voltage and current noise sources of $V_n(f) = 20\ \text{nV}/\sqrt{\text{Hz}}$ and $I_n(f) = 10\ \text{pA}/\sqrt{\text{Hz}}$, respectively. Estimate the total rms noise voltage at V_o.

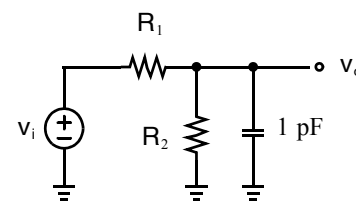
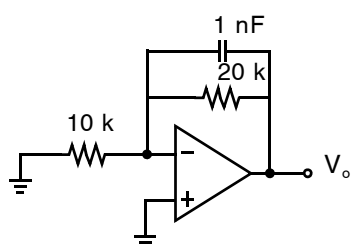


Fig. P9.16



Opamp

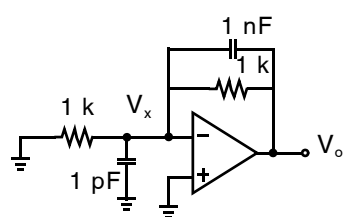
$V_n(f) = 20\ \text{nV}/\sqrt{\text{Hz}}$

$I_{n+}(f) = I_{n-}(f) = 10\ \text{pA}/\sqrt{\text{Hz}}$

$f_t = 1\ \text{MHz}$

Fig. P9.17

9.18 Assuming an ideal noiseless opamp, find the noise level at V_x and V_o for the opamp circuit shown in Fig. P9.18. By what factor is the noise value at V_o larger (or smaller) than $kT/1\ \text{nF}$? How do you account for this increase (or decrease)? Also explain why the noise value at V_x is smaller than $kT/1\ \text{nF}$.



Ideal noiseless opamp

Fig. P9.18

9.19 Consider an inductor of value L and an arbitrary resistor in parallel, as shown in Fig. P9.19. Show that the current noise, $i_{no}(t)$, has a noise value given by

$$I^2_{no(rms)} = \frac{kT}{L}$$

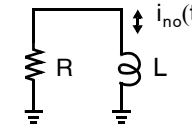


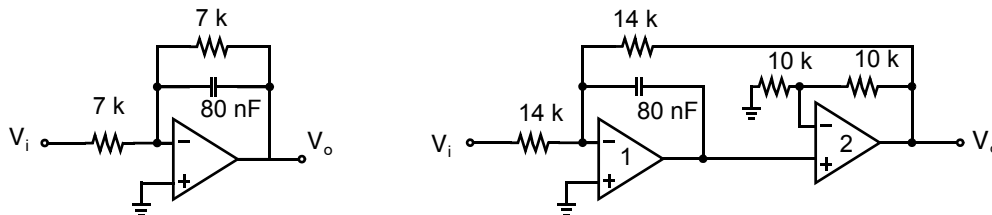
Fig. P9.19

9.20 The two circuits shown in Fig. P9.20 realize a first-order, low-pass transfer function from V_i to V_o. Assume the opamps are ideal and noiseless.

a. Show that the two circuits have the same input–output transfer function.

b. Estimate the total output noise for circuit I, using only dominant noise sources.

c. Repeat (*b*) for circuit II.

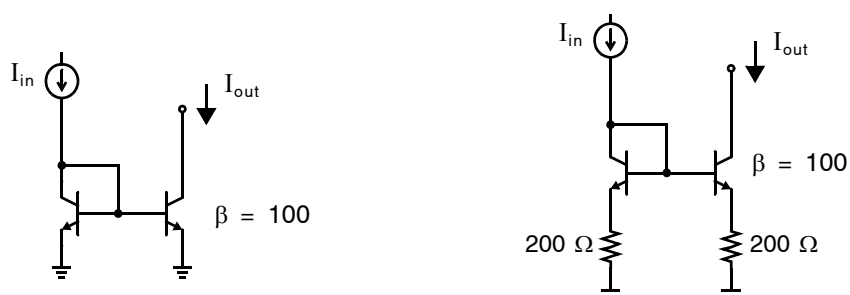


I II

Fig. P9.20

9.21 Modify circuit I in Fig. P9.20 such that the new circuit has the same transfer function but uses an 80-pF capacitor instead of an 80-nF capacitor. If the opamp is ideal and noiseless, what is the new total output noise?

9.22 Consider the two bipolar current mirrors shown in Fig. P9.22, where I_{in} is a 1-mA bias current plus a 100-μA(rms) signal current. Assuming that the base resistance for each transistor is $r_b = 330\ \Omega$ and dominates the output noise, estimate the resulting SNR (in dB) for the two current mirrors over a 50-MHz bandwidth (also assume the output noise is white).



(*a*) (*b*)

Fig. P9.22

9.23 Consider the NMOS current mirror depicted in Fig. P9.23. You may assume that Q_2 is in active mode, and that the load connected to I_{out} has a much lower impedance than the mirror's output impedance. Find an expression for the current noise spectral density that appears superimposed upon the dc current I_{out} in terms of the value of C and all device and small-signal model parameters for Q_1 and Q_2. (You may neglect all other parasitic capacitors in the circuit.)

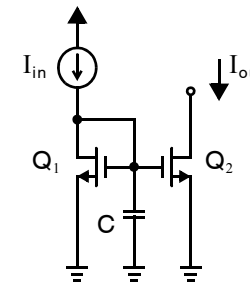


Fig. P9.23

9.24 In the common-source amplifier of Fig. P9.24, $I_{in} = 1$ mA and Q_1 is in active mode with $(W/L) = (30\ \mu m/0.2\ \mu m)$. Using the device parameters for the 0.18-μm technology in Table 1.5 and assuming $\gamma = 2/3$:

a. Sketch the voltage noise spectral density at v_o.

b. Find the total rms voltage noise at v_o.

9.25 Repeat Problem 9.24 but this time use device parameters for the 45 nm technology in Table 1.5 and assume $\gamma = 2$ with $(W/L) = (7.5\ \mu m/0.05\ \mu m)$.

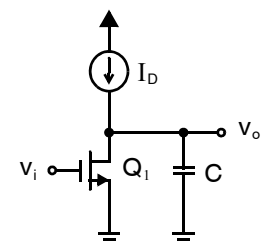


Fig. P9.24

9.26 Select W/L and C_L in Fig. P9.26 so that the circuit has a 3-dB bandwidth of 50 MHz and a SNR of 45 dB is observed at v_{out} when $v_{in} = 100\ \text{mV} + (100\ \text{mV})\sin(2\pi \cdot 10^3 t)$. Use the device parameters for the 0.18-μm technology in Table 1.5 and assume $\gamma = 2/3$.

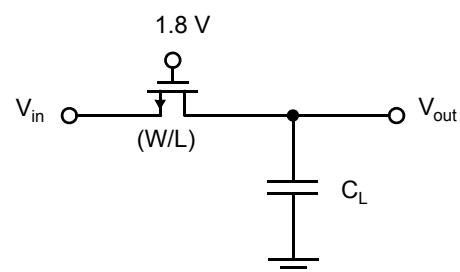


Fig. P9.26

9.8.3 Section 9.4: Noise Analysis Examples

9.27 Estimate the total output noise rms value for a low-pass filter, as shown in Fig. 9.18, when $C_f = 1$ nF, $R_f = 16$ k, $R_1 = 1.6$ k, and $R_2 = 0$. Also, find the SNR for an input signal equal to 100 mV rms. Assume that the noise voltage of the opamp is given by $V_n(f) = 20\ \text{nV}/\sqrt{\text{Hz}}$, that both its noise currents are $I_n(f) = 0.6\ \text{pA}/\sqrt{\text{Hz}}$, and that its unity-gain frequency equals 2 MHz.

9.28 Consider the CMOS differential input stage shown in Fig. 9.20, where Q_5 supplies a bias current of 100 μA, resulting in $g_{m1} = g_{m2} = 1$ mA/V and $g_{m3} = g_{m4} = 0.5$ mA/V. Find the equivalent input noise spectral density associated with thermal noise. If the bias current is doubled, how does the equivalent input noise density change?

9.8.4 Section 9.5: Dynamic Range Performance

9.29 Calculate the percentage THD value of a signal in which the fundamental component is at 1 MHz and has a voltage level of 1 V_{rms}, and the only significant harmonics are the first three above the fundamental. These harmonics, at 2 MHz, 3 MHz, and 4 MHz, are measured to have voltage levels of 1 mV_{rms}, 0.5 mV_{rms}, and 0.3 mV_{rms}, respectively.

9.30 If a circuit is measured to have $IIP_3 = 10$ dBm and has a gain of 6 dB, what output-signal level should be used such that the third-order intermodulation products are 60 dB below the fundamental?

9.31 At an input-signal level of -4 dBm, an intermodulation ratio of -40 dB was measured in a filter with a gain of 6 dB. Calculate the value of the input and output third-order intercept points. What input-signal level should be applied if we want an intermodulation ratio of -50 dB? If the noise power at the output is measured to be -60 dBm, what is the expected SFDR and what output-signal level does it correspond to?

9.32 An analog circuit has an input-output relationship governed by

$$v_{out} = 10v_{in} + 0.02v_{in}^3$$

The circuit has a total output noise of $N_o = 20\ \mu V_{rms}$. Find $SNDR_{max}$. At what input amplitude is $SNDR_{max}$ achieved?

CHAPTER

10 Comparators

Perhaps the second most widely used electronic components (after amplifiers) are comparators. A comparator is used to detect whether a signal is greater or smaller than zero, or to compare the size of one signal to another. As we will see in Chapter 17, comparators are used in large abundance in A/D converters. They also find widespread use in many other applications, such as data transmission, switching power regulators, and others. In this chapter, we look at comparator design and practical limitations, where a number of different approaches are discussed. First, we examine a simplistic approach of using an open-loop opamp for a comparator. Although this approach is too slow for practical applications, it is a good example to use when we discuss several design principles for minimizing input-offset voltage and charge-injection errors. A number of other approaches are also described, such as multiple-stage comparators, positive-feedback track-and-latch comparators, and fully differential comparators. Finally, examples of both CMOS and bipolar comparator circuits are also described.

10.1 COMPARATOR SPECIFICATIONS

10.1.1 Input Offset and Noise

The input offset of a comparator is the input voltage at which its output changes from one logic level to the other (neglecting noise). In an ideal comparator, this value would be zero. In a practical comparator, input offset might be caused by device mismatches (Chapter 2) or might be inherent in the design of the comparator.

Key Point: *The input offset of a comparator is the input at which the output is equally likely to be high or low. Noise determines how the output statistics vary as a function of input.*

Random circuit noise can cause the output of a comparator to change from one logic state to the other, even when the comparator input is held constant. Hence, in order to measure the input-offset voltage in the presence of circuit noise, one would look for the input voltage that results in the output state of the comparator being high and low with equal likelihood. The input-referred noise is then observed by changing the input around this value and observing how the output statistics vary as a function of the input voltage around this dc offset.

EXAMPLE 10.1

The output of a comparator is observed while slowly sweeping the input over the range −10 mV to +10 mV. The fraction of time that the comparator output is high is plotted as a function of the input voltage in

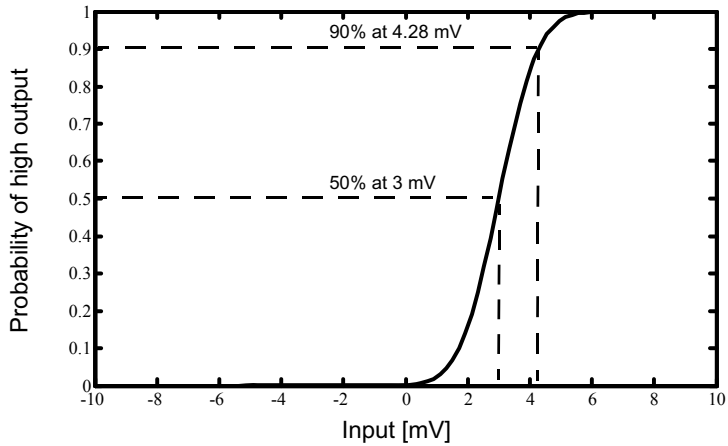


Fig. 10.1 Probability of a high output logic level with varying input level for a particular comparator.

Fig. 10.1. Estimate the comparator's input offset and noise. For this comparator, what is the smallest input that can be resolved with an error rate of 10^{-12}?

Solution

The plot in Fig. 10.1 is recognized as a normal distribution. At an input of 3 mV, the comparator output is equally likely to be high or low; hence, the comparator's input offset is 3 mV. The 90% confidence interval for a normal distribution is at 1.2816 times its standard deviation (rms) value. In Fig. 10.1, this occurs at an input of 4.28 mV, implying an rms input noise of

$$\frac{4.28 - 3}{1.2816}\ \text{mV} = 1\ \text{mV}_{\text{rms}}$$

For an error rate of 10^{-12}, the input must exceed the input offset by 7.0344 times the rms noise level. Hence, for a high output logic level the required input voltage is $3 + 7.0344 \cdot 1\ \text{mV} = 10.0\ \text{mV}$. For a low output logic level, the input must be below –4 mV.

10.1.2 Hysteresis

Key Point: *Hysteresis is "memory" in a comparator that tends to make it remain in its past output state, usually due to charge retained on internal circuit capacitances.*

Many comparator designs have an input threshold that is dependent upon the past state of the comparator. Specifically, most comparators have a tendency to remain in their past state. For example, when a comparator's output is at a high logic level, it might remain so for input voltages as low as –20 mV, but when that same comparator's output is at a low logic level an input voltage of +20 mV might be required to cause the input to change to a high logic level. The comparator's "memory" results from charge stored on the its internal capacitances, and is referred to as *hysteresis*. Generally, hysteresis is a nonideality in comparators that should be minimized, usually by ensuring all internal capacitances are completely discharged periodically. However, in some applications hysteresis may be used

to advantage. For example, if the comparator is to be operated with near-zero input for long periods of time, some hysteresis will prevent the comparator's output from toggling back and forth between high and low logic levels due to noise.

10.2 USING AN OPAMP FOR A COMPARATOR

A simplistic approach for realizing a comparator is to use an open-loop opamp, as shown in Fig. 10.2. The main drawback to this approach is the slow response time since the opamp output has to slew a large amount of output voltage and settles too slowly. However, temporarily ignoring this slow response time, we will first investigate its input offset voltage.

The simplistic opamp approach shown in Fig. 10.2 has a resolution limited to the input-offset voltage of the opamp. This offset might be on the order of 2 mV to 5 mV for typical MOS processes, which is inadequate for many applications. An alternative architecture that can resolve signals with accuracies much less than the input-offset voltages of opamps is shown in Fig. 10.3 [McCreary, 1975; Yee, 1978]. Although this circuit has been used many times in early analog-to-digital converters, as we will see, it is not preferable nowadays. However, it is a simple example that can be used to illustrate many important design principles. The circuit of Fig. 10.3 operates as follows: During ϕ_1, known as the *reset phase,* the bottom plate[1] of the capacitor C (i.e., the left side of capacitor C) is connected to ground, and the top plate is connected to the inverting input of the opamp. At the same time, the output of the opamp is also connected to the inverting input of the opamp by closing switch S_r. Assuming the opamp is ideal, this connection causes the capacitor to be charged to zero volts. Next, during the *comparison phase*, the reset switch, S_r, is turned off, and the bottom plate of the capacitor is connected to the input voltage. The opamp is now in an open-loop configuration. If the input signal is greater than zero, the output of the opamp swings to a large negative voltage. If the input signal is less than zero, the output of the opamp swings to a large positive voltage. These two cases are easily resolved and the decision can be stored using a simple digital latch.

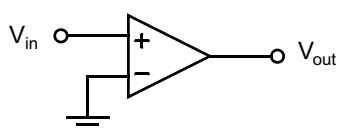


Fig. 10.2 A simplistic approach of using an open-loop opamp for a comparator.

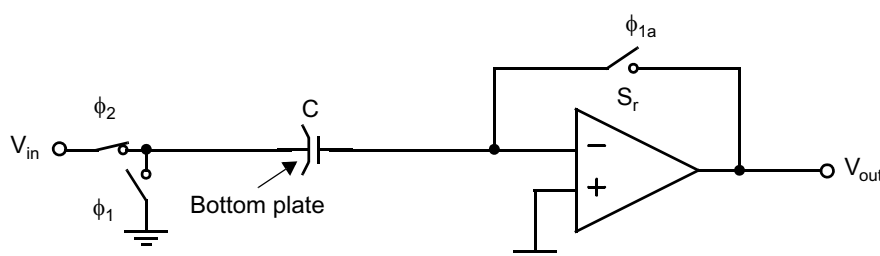


Fig. 10.3 Cancelling the offset voltage of a comparator—the comparator here must be stable, with unity-gain feedback during ϕ_{1a}. (ϕ_{1a} is a slightly advanced version of ϕ_1 so that charge-injection effects are reduced.)

1. The bottom plate of a capacitor has significant parasitic capacitance between it and ground. Therefore, it is almost always connected to the less sensitive input side rather than to the critical amplifier side (see Chapter 14).

The limitations of this approach become apparent when one considers nonideal opamps, which have finite gains and require compensation to be stable during the reset phase.

EXAMPLE 10.2

Consider the case in which a 0.2-mV signal must be resolved using the circuit shown in Fig. 10.3, where the opamp's output should be 2 V. Assuming the opamp's unity-gain frequency is 10 MHz, find the maximum clocking rate of the comparator circuit if reset and comparison phases are equal and if six time constants are allowed for settling.

Solution

After the comparison phase, the output of the opamp should have a 2-V difference between the cases in which the input signal is either –0.1 mV or +0.1 mV.[2] As a result, the opamp gain must be at least 10,000 V/V. By assuming that the dominant-pole compensation is used to guarantee stability during the reset phase, we obtain an open-loop transfer function for the opamp similar to that shown in Fig. 10.4. Here, the –3-dB frequency of the opamp is given by

$$f_{-3\ dB} = \frac{f_t}{A_0} = 1\ \text{kHz} \tag{10.1}$$

During the comparison phase, the output of the opamp will have a transient response similar to that of a first-order system that has a time constant given by

$$\tau = \frac{1}{2\pi f_{-3\ dB}} \cong 0.16\ \text{ms} \tag{10.2}$$

If six time constants are allowed for settling during the comparison phase, then approximately 1 ms is needed for the comparison phase. Assuming the reset time is the same as the comparison time, the clock frequency can be no greater than 500 Hz—a frequency that is much too slow for most applications.

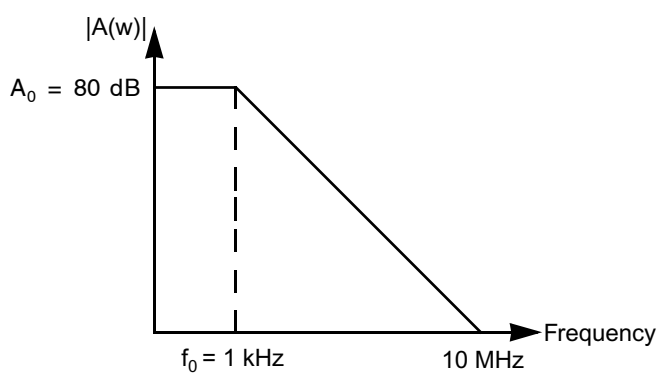


Fig. 10.4 The open-loop transfer function of the opamp used to realize the comparator.

2. This assumes that ±1-V power supplies are being used and that digital output values are the same as the power-supply voltages, as is typical for CMOS. If a single power supply voltage of 2 V is used, then analog ground can be considered to be at 1 V.

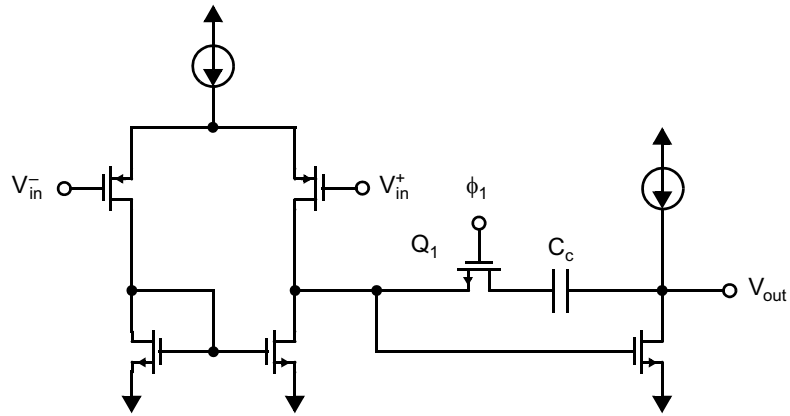


Fig. 10.5 An opamp that has its compensation capacitor disconnected during the comparison phase.

One possibility for speeding up the comparison time is to disconnect the compensation capacitor during the comparison phase. For example, a simplified opamp schematic is shown in Fig. 10.5. In this opamp, transistor Q_1 is used to achieve lead compensation when it is on during the reset phase, ϕ_1. During the comparison phase, Q_1 is turned off, which disconnects the compensation capacitor, C_C, thereby greatly speeding up the opamp during this phase. Using this technique, it is possible to use clock frequencies ten to fifty times greater than would otherwise be the case—perhaps as high as 25 or 50 kHz in our example. If this is adequate, then the approach is reasonable. Unfortunately, this often isn't adequate, and other approaches, to be described shortly, are necessary.

Key Point: *Regular opamps are compensated to ensure stability in closed-loop configurations, resulting in a low 3-dB bandwidth and, hence, very slow settling if used open-loop as a comparator.*

One superior aspect of the approach just described is that the input capacitor, C in Fig. 10.3, is never charged or discharged during operation. Specifically, in phase ϕ_1, C is always charged to 0 V. In phase ϕ_2, the top plate, connected to the inverting input of the opamp, is open-circuited at that time (assuming the parasitic capacitors are ignored), and the voltage across capacitor C remains at 0 V (in other words, the top-plate voltage follows V_{in}). This approach greatly minimizes the charge required from the input when V_{in} changes. If the switches attached to the bottom plate had their phases interchanged, then the comparison operation would be noninverting, but now capacitor C must be charged or discharged during the reset phase since the bottom plate follows V_{in}, whereas the top plate remains at virtual ground. Normally, we like to use a reasonably large input capacitor to minimize clock-feedthrough effects (described in Section 10.3). This charging/discharging requirement puts severe constraints on the input signal source, and, thus, the clock phasing shown in Fig. 10.3 should be used. Besides, it is usually possible to tolerate an inverting comparison since it can be made noninverting through the use of a digital inversion somewhere else in the system.

10.2.1 Input-Offset Voltage Errors

In switched-capacitor comparators, such as that shown in Fig. 10.3, input offset is not a problem since it is stored across the capacitor during the reset phase, and then the error is cancelled during the comparison phase. To appreciate this cancellation, assume the opamp has an input-offset voltage error that is modelled as a voltage source in series with one of the opamp's inputs. The circuit configuration during the reset phase is shown in

Key Point: *An input capacitor is sometimes used to sample the input offset voltage of a comparator during a reset phase, and then subtract it from the input during the comparison phase.*

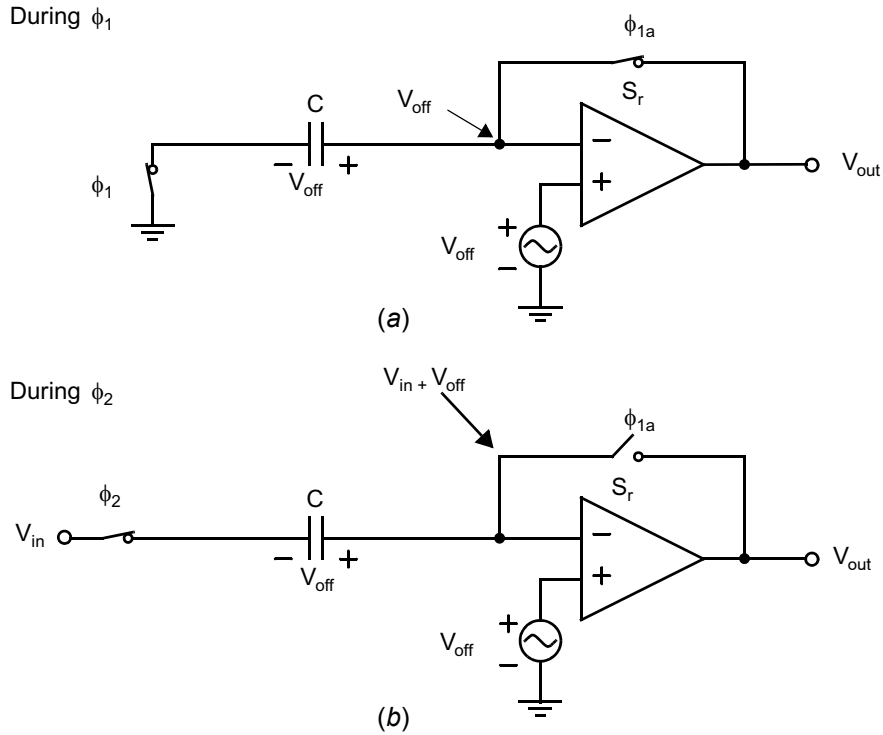


Fig. 10.6 The circuit configuration (*a*) during the reset phase, and (*b*) during the comparison phase, assuming the opamp has an input-offset voltage given by V_{off}.

Fig. 10.6(*a*). Assuming the opamp gain is very large, then the inverting input of the opamp is at the voltage V_{off}, which implies that the input capacitor is charged to V_{off} during this phase. Next, during ϕ_2, as shown in Fig. 10.6(*b*), the left side of capacitor C is connected to the input voltage. The right side of the capacitor has a voltage given by $V_{in} + V_{off}$, which results in the comparator output becoming negative if V_{in} is greater than zero or positive if V_{in} is less than zero, regardless of the value of V_{off}. Furthermore, notice that very low-frequency input-referred noise is indistinguishable from input-offset, and hence is cancelled along with V_{off}. Therefore, not only does this technique eliminate input-offset voltage errors, but it also minimizes errors caused by low-frequency 1/f noise, which can be large in CMOS microcircuits.

10.3 CHARGE-INJECTION ERRORS

Perhaps the major limitation on the resolution of comparators is due to what is referred to as *charge-injection,* also commonly called *clock feedthrough*. This error is due to unwanted charges being injected into the circuit when the transistors turn off. For the comparator in Fig. 10.3, the switches are normally realized by either n-channel transistors alone or CMOS transmission gates (which consist of n -channel transistors in parallel with p-channel transistors, both of which must turn off). When MOS switches are on, they operate in the triode region and have zero volts between their drain and their source. When MOS switches turn off, charge errors occur by two mechanisms. The first is due to the channel charge, which must flow out from the channel region of the transistor

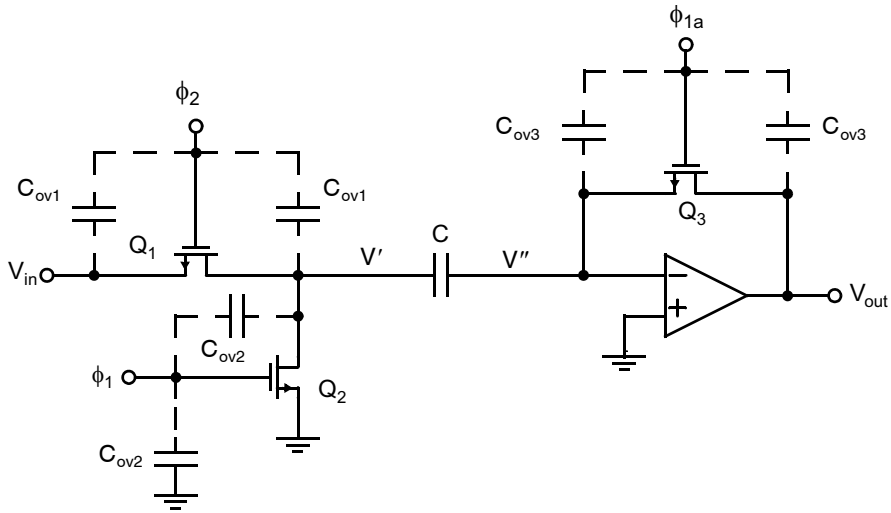


Fig. 10.7 The comparator in Fig. 10.3, with n-channel switches and overlap parasitic capacitances shown. Transistors turned on also have channel charge.

to the drain and the source junctions.[3] The channel charge of a transistor that has zero V_{DS} is given by

$$Q_{ch} = WLC_{ox}V_{eff} = WLC_{ox}(V_{GS} - V_t) \tag{10.3}$$

This charge often dominates. The second charge (typically smaller, unless V_{eff} is very small) is due to the overlap capacitance between the gate and the junctions.

Figure 10.7 shows the comparator of Fig. 10.3, where the switches have been realized as n-channel transistors. Also shown in Fig. 10.7 are the parasitic capacitances due to gate-drain and gate-source overlap capacitors. Finally, although it is not shown in Fig. 10.7, one should remember to include the channel charge dispersion of any transistors that change from being on to off.[4] In the circuit shown, all transistors are n-channel, implying that the channel charge is negative.

Consider first when Q_3 turns off. If the clock waveform is very fast, the channel charge due to Q_3 will flow equally out through both junctions [Shieh, 1987]. The $Q_{ch}/2$ charge that goes to the output node of the opamp will have very little effect other than causing a temporary glitch. However, the $Q_{ch}/2$ charge that goes to the inverting-input node of the opamp will cause the voltage across C to change, which introduces an error. Since this charge is negative, for an n-channel transistor, the node voltage V″ will become negative. The voltage change due to the channel charge is given by

$$\Delta V'' = \frac{(Q_{ch}/2)}{C} = -\frac{V_{eff3}C_{ox}W_3L_3}{2C} = -\frac{(V_{DD} - V_{tn})C_{ox}W_3L_3}{2C} \tag{10.4}$$

since the effective gate-source voltage of Q_3 is given by $V_{eff3} = V_{GS3} - V_{tn} = V_{DD} - V_{tn}$. The preceding voltage change in V″ is based on the assumption that Q_2 turns off slightly after Q_3 turns off. More will be said about this assumption shortly.

3. For a transistor in the triode region that has zero V_{DS}, it doesn't matter which junctions are called the drain and the source, since both junctions are at the same potential.

4. Transistors also accumulate channel charge when turning on, but this does not typically affect circuit performance since the charge comes from low-impedance nodes.

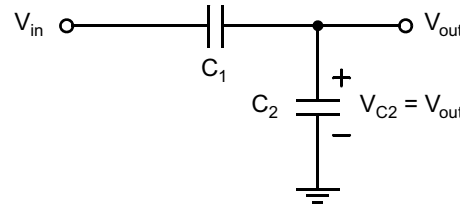


Fig. 10.8 A capacitor divider.

Key Point: *As a MOSFET goes from triode mode to off, the charge stored in the channel region must flow out of the transistor terminals; in a comparator, the charge may accumulate on circuit capacitances causing offset.*

To calculate the change in voltage due to the overlap capacitance, it is first necessary to introduce the capacitor-divider formula. This formula is used to calculate the voltage change at the internal node of two series capacitors, when the voltage at one of the end terminals changes. This situation is shown in Fig. 10.8, where it is assumed that V_{in} is changing and we want to calculate the change in $V_{out} = V_{C2}$. The series combination of C_1 and C_2 is equal to a single capacitor, C_{eq}, given by

$$C_{eq} = \frac{C_1 C_2}{C_1 + C_2} \tag{10.5}$$

When V_{in} changes, the charge flow into this equivalent capacitor is given by

$$\Delta Q_{eq} = \Delta V_{in} C_{eq} = \Delta V_{in} \frac{C_1 C_2}{C_1 + C_2} \tag{10.6}$$

All of the charge that flows into C_{eq} is equal to the charge flow into C_1, which is also equal to the charge flow into C_2. Thus, we have

$$\Delta V_{out} = \Delta V_{C2} = \frac{\Delta Q_{C2}}{C_2} = \frac{\Delta V_{in} C_1}{C_1 + C_2} \tag{10.7}$$

This formula is often useful when calculating charge flow in integrated circuits. It can be applied to the circuit of Fig. 10.7 to calculate the change in V'' due to the overlap capacitance of Q_3 when it turns off. For this case, we have $C_1 = C_{ov}$, $C_2 = C$, and $\Delta V_{in} = -(V_{DD} - V_{SS})$. This assumes the clock signals change from V_{DD} to V_{SS}. The change in V'' due to the overlap capacitance is now found to be given by

$$\Delta V'' = \frac{-(V_{DD} - V_{SS}) C_{ov}}{C_{ov} + C} \tag{10.8}$$

This change is normally less than that due to the change caused by the channel charge since C_{ov} is small.

EXAMPLE 10.3

Assume that a resolution of ±2.5 mV is required from the circuit of Fig. 10.7. The following values are given: $C = 0.1$ pF, $C_{ox} = 8.5$ fF/(μm)^2, $(W/L)_3 = 4\ \mu m/0.2\ \mu m$, $L_{ov} = 0.04\ \mu m$, $V_{tn} = 0.45$ V, and $V_{DD} = 1$V. What is the change in $\Delta V''$ when Q_3 turns off?

Solution

Using (10.4), we have $\Delta V'' = -53$ mV. The overlap capacitance is given by $C_{ov} = W_3 L_{ov} C_{ox} = 1.4$ fF. Using (10.8), this gives a voltage change in $\Delta V''$ of $\Delta V'' = -2.8$ mV, which is smaller than the change due to the channel charge, but not insignificant. The total change is found by adding the two effects so that $\Delta V'' = -56$ mV.

This result is on the order of 10 times greater than the value of ±2.5 mV that should be resolved. Clearly, additional measures should be taken to minimize the effects of charge-injection if this resolution is required.

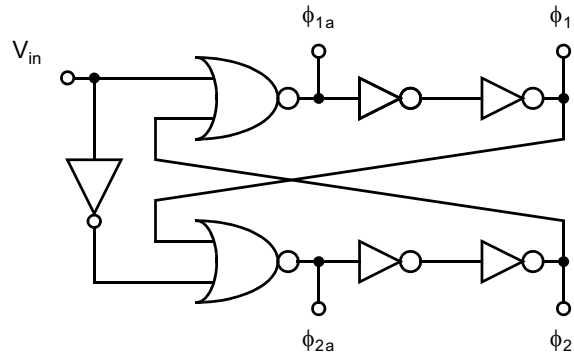


Fig. 10.9 A clock generator suitable for generating the desired clock waveforms for the comparator in Fig. 10.3.

10.3.1 Making Charge-Injection Signal Independent

> ***Key Point:*** *The deleterious effects of charge-injection can be reduced by using "advanced" clocks which open-circuit the discharge paths through which signal-dependent charge-injection flows.*

The charge-injection due to transistors Q_1 and Q_2 may cause temporary glitches, but it will have much less effect than the charge-injection charge-injection due to Q_3 if we assume that Q_2 turns off slightly after Q_3 does [Haigh, 1983]. This time difference is the reason why the clock voltage of Q_2 is denoted ϕ_1, whereas the clock voltage of Q_3 is denoted ϕ_{1a} (ϕ_1 advanced). The argument for this arrangement is as follows: When Q_2 turns off, its charge-injection causes a negative glitch at V', but this will not cause any change in the charge stored in C since the right side of C is connected to an effective open circuit, assuming that Q_3 has already turned off. Later, when Q_1 turns on, the voltage V' will settle to V_{in} regardless of the previous charge-injection of Q_2. Thus, the voltage at the left of C is unaffected by the charge-injection of Q_2, and the voltage across C is also unaffected by the charge-injection of Q_2. Therefore, the voltage V'' is unaffected by the charge-injection of Q_2. This is not the case if Q_2 is turned off at the same time or before Q_3 is turned off. The charge-injection of Q_1 has no effect for a simpler reason—the comparison has already been made when it turns off. In addition, its charge-injection has no effect when it turns on because the right side of C is connected to an open circuit at that time, assuming the clocks do not overlap. *In summary, by turning off ϕ_{1a} first, the cir cuit is affec ted only by the char ge-injection of Q_3 and not by any of the othe r switches.*

Figure 10.9 shows a simple digital circuit [Martin, 1980] that is capable of generating the desired clock waveforms. The waveforms ϕ_1 and ϕ_{1a} do not overlap with ϕ_2 and ϕ_{2a}. Also, ϕ_{1a} will be advanced slightly (by two inverter delays), compared to ϕ_1.

10.3.2 Minimizing Errors Due to Charge-Injection

The simplest way to reduce errors due to charge-injection is to use larger capacitors. For our previous example, capacitors on the order of about 10 pF guarantee that the clock-feedthrough errors are approximately 0.5 mV. Unfortunately, this large amount of capacitance would require a large amount of silicon area. Also, integrated capacitors have parasitic capacitances between their terminals and the substrate that might be about 20 percent of the size of the realized capacitor. Thus, a parasitic capacitance of about 2 pF would have to be driven by the input circuits, which would greatly slow down the circuits.

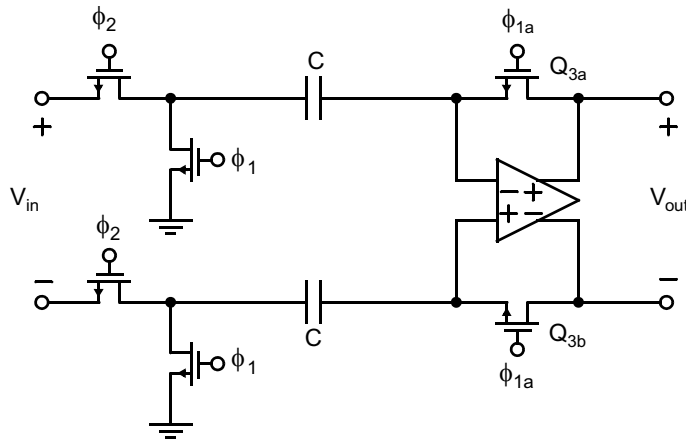


Fig. 10.10 A fully differential, single-stage, switched-capacitor comparator.

An alternative approach for minimizing errors due to charge-injection is to use fully differential design techniques for comparators, similar to what is often done for opamps. A simple example of a one-stage, switched-capacitor, fully differential comparator is shown in Fig. 10.10. Note that the ground connections to the reset switches φ_1 in Fig. 10.10 may be replaced by a common-mode voltage reference located midway between the analog supplies. Ideally, when the comparator is taken out of reset mode, the clock feedthrough of reset switch Q_{3a} matches the clock feedthrough of Q_{3b}. In this case, the common-mode voltage may be slightly affected, but the differential input voltage is unaffected. The only errors now are due to mismatches in the clock feedthrough of the two switches, which will typically be at least ten times smaller than in the single-ended case. For this reason, virtually all modern integrated comparators utilize fully differential design techniques.

A third alternative that can be used along with fully differential design techniques is to realize a multi-stage comparator [Poujois, 1978; Vittoz, 1985], where the clock feedthrough of the first stage is stored on coupling capacitors between the first and second stage, thereby eliminating its effect. The clock feedthrough of the second stage can be stored on the coupling capacitors between the second and the third stages, and so on. Although this technique is almost always used along with fully differential configurations, we will explain the technique using single-ended configurations, for simplicity. Consider the three-stage comparator shown in Fig. 10.11(*a*), along with the appropriate clock waveforms shown in Fig. 10.11(*b*). Consider the time when ϕ_1' drops and the switch of the first stage has charge-injection. Figure 10.12 illustrates this case, along with the parasitic capacitances of the first-stage switch. When ϕ_1' drops, Q_1 injects charge into both the inverting input and the output of the first stage through parasitic capacitors C_{p1} and C_{p2}, respectively. The charge injected at the first stage output causes only a temporary glitch there. The charge injected at the inverting input causes this node to become negative. The amount by which this node voltage becomes negative is calculated using the analysis method described previously. Let us assume this amount will be in the range of tens of millivolts, as calculated in Example 10.3. After the inverting input becomes negative, the output of the first stage becomes positive by an amount equal to the negative transition of the inverting input multiplied by the first stage's gain. However, at this time, ϕ_1'' is still high. Therefore, the second stage is still being reset and C_2 is charged up to the output error caused by the clock feedthrough of the first stage, thereby eliminating its effect. In a similar manner, when ϕ_1'' turns off and the second stage goes from closed-loop reset mode to open-loop comparison mode, the third stage is still in reset mode and the clock feedthrough of the second stage is stored on coupling capacitor C_3. Finally, when the third stage turns off, its charge-injection is not cancelled. However, the error it causes in resolving an input voltage to all three stages is small, equal to the voltage transition at the inverting input of the third stage divided by the negative of the gains of the first two stages (i.e., this is the input voltage needed to cancel the effect of the clock feedthrough of the third stage coming out of reset mode).

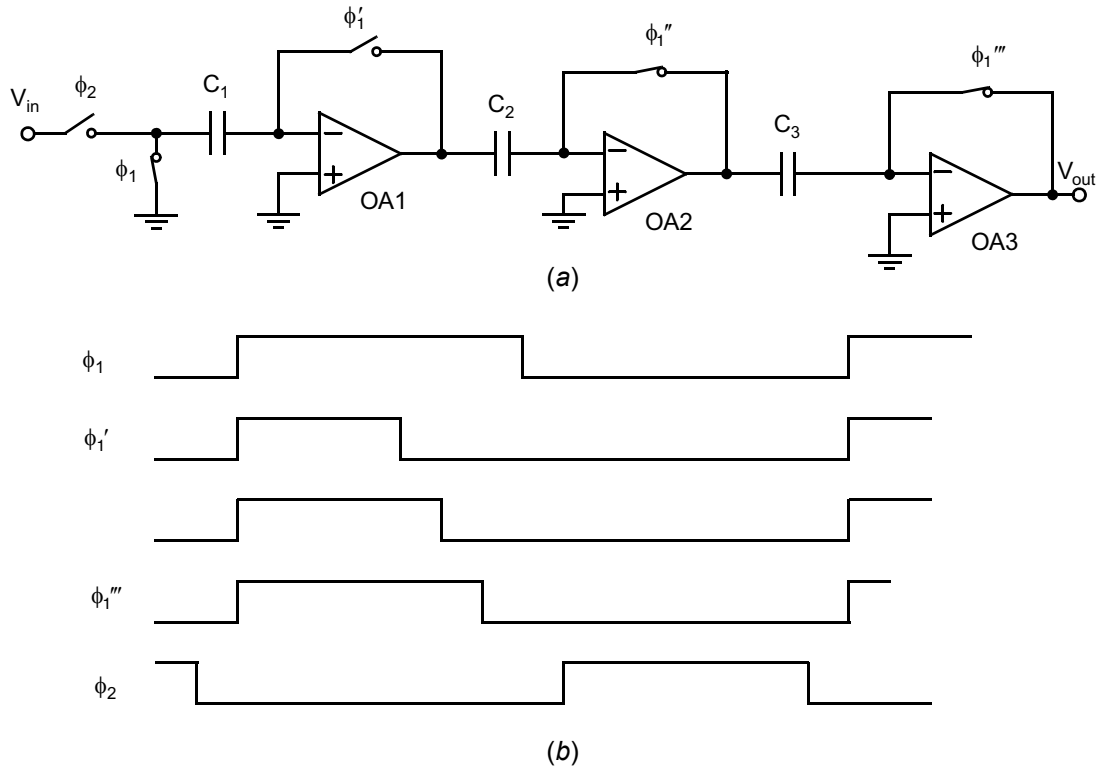


Fig. 10.11 (*a*) A multi-stage comparator used to eliminate clock feedthrough errors, with (*b*) the appropriate clock waveforms.

In a variation on this multi-stage approach [Poujois, 1978], the first stage was not reset. However, the input-offset voltage of the first stage was still cancelled by storing it on coupling capacitors between the first and second stage during a reset mode. This approach is described Section 10.5.

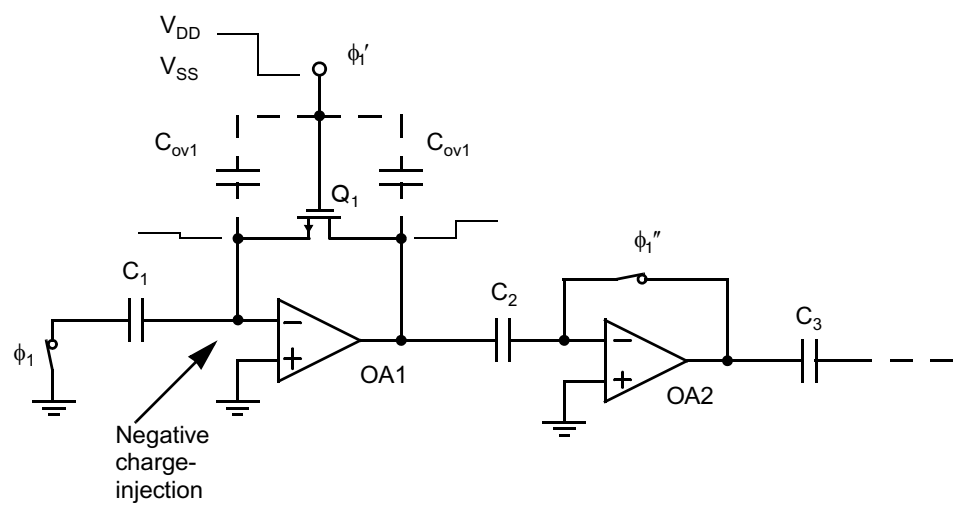


Fig. 10.12 The first stage of the comparator in Fig. 10.11, when the first stage is injecting charge.

EXAMPLE 10.4

Assume all capacitors are 0.1 pF, transistor sizes are $W/L = 4\ \mu m/0.2\ \mu m$, $C_{ox} = 8.5\ fF/(\mu m)^2$, $L_{ov} = 0.04\ \mu m$, $V_{tn} = 0.45V$, and $V_{DD} = 2\ V$, and each stage has a gain of 20. What is the input-offset voltage error caused by the clock feedthrough of the third stage when the third stage comes out of reset?

Solution

The values used are identical to those used in Example 10.3. Therefore, the charge-injection at the inverting input of the third stage is the same as that found in the solution to Example 10.3, or −56 mV. The input signal that can overcome this value 56 mV is given by

$$\Delta V_{in} = \frac{56\ mV}{A_1 A_2} = 140\ \mu V \tag{10.9}$$

Thus, the equivalent input-offset voltage is 140 μV, which is much better than the resolution found in Example 10.3 of 56 mV.

10.3.3 Speed of Multi-Stage Comparators

The approach just described can be used to realize very-high-resolution comparators, especially when it is combined with fully differential circuit design techniques. However, this approach does have the limitation that it requires multiple-phase clock waveforms, which slow down the circuits.

Although the multi-stage approach is limited in speed due to the need for the signal to propagate through all the stages, it can still be reasonably fast because each of the individual stages can be made to operate fast. Typically, each stage consists of a single-stage amplifier that has only a 90° phase shift and therefore does not need compensation capacitors (i.e., each stage has a 90° phase margin without compensation). To see why this multi-stage approach can be fast, consider the simplified case of a cascade of first-order, uncompensated inverters, as shown in Fig. 10.13. This circuit is a very rough approximation of a multi-stage comparator when it is not being reset. The parasitic load capacitance at the output of the ith stage is approximately given by

$$C_{pi} \cong C_{0\text{-}i} + C_{gs\text{-}i+1} \tag{10.10}$$

where $C_{0\text{-}i}$ is the output capacitance of the $_i$th stage, which would normally be due to junction capacitances, and $C_{gs\text{-}i+1}$ is the gate-source capacitance of the input transistor of the succeeding stage. Equation (10.10) will not necessarily be true for the last stage, but the load of the last stage should be of roughly the same magnitude. If one

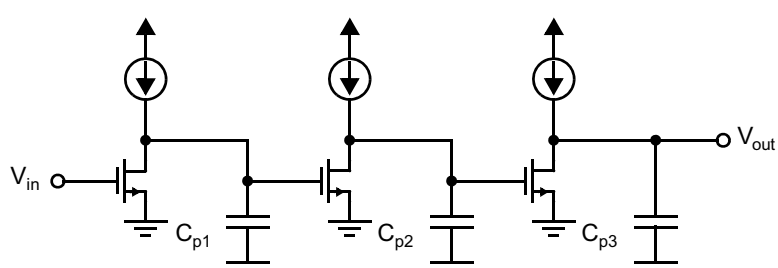


Fig. 10.13 Realizing a comparator by using a cascade of first-order gain stages.

assumes the stages are matched, then normally $C_{0\text{-}i} < C_{gs\text{-}i+1}$ since junction capacitances are usually less than gate-source capacitances, implying that one can usually assume

$$C_{pi} < 2C_{gs\text{-}i} \tag{10.11}$$

The unity-gain frequency of a single stage is then on the order of

$$\omega_{ti} \sim \frac{g_{mi}}{2C_{gs\text{-}i}} \tag{10.12}$$

or larger, where g_{mi} is the transconductance of the input capacitor of the i th stage. Thus, the unity-gain frequency of a single gain-stage is on the order of one-half the unity-gain frequency of a single transistor. If one assumes the ith stage has a dc gain $A_{0\text{-}i}$ and that the ith stage is well described by a first-order transfer function, then the transfer function of a single stage is approximately given by

$$A_i(s) = \frac{A_{0\text{-}i}}{1 + s/\omega_{p\text{-}i}} \tag{10.13}$$

where

$$\omega_{p\text{-}i} \cong \frac{\omega_{t\text{-}i}}{A_{0\text{-}i}} \sim \frac{g_{mi}}{2A_{0\text{-}i}C_{gs\text{-}i}} \tag{10.14}$$

Thus, the –3-dB frequency of a single stage is on the order of one-half the unity-gain frequency of a transistor divided by the gain of the stage. If one has a cascade of n stages, then the overall transfer function is given by

$$A_{total}(s) = \prod A_i(s) \tag{10.15}$$

This result can be approximated by a first-order transfer function, where all higher-order terms are ignored, resulting in

$$A_{total}(s) \cong \frac{\prod A_{0\text{-}i}}{1 + s\sum 1/\omega_{p\text{-}i}} \cong \frac{A_0^n}{1 + ns/\omega_{p\text{-}i}} \tag{10.16}$$

Thus, a cascade of n stages has a time constant approximately given by

$$\tau_{total} \cong \frac{2nA_0C_{gs}}{g_m} \cong 2nA_0\tau_T \tag{10.17}$$

where $\tau_T = C_{gs}/g_m$ is the approximate transit time of a single transistor (i.e., 1 over the unity-gain frequency of a single transistor). In other words, the time constant of the cascade of first-order stages is approximately equal to n times the time constant of a single stage, which is roughly given by $2A_0$ times the transit time of a single transistor. This result should be compared to using a single opamp, which, for the same overall gain, will have a time constant considerably greater than $2A_0^n$ times the transit time of a single transistor. Although the preceding speed estimate is valid for simple single-ended stages only, the general principles apply to more complicated, fully differential stages as long as each stage is first order and doesn't require compensation.

Equation (10.17) can be simplified further. Recall from Chapter 1 that

$$C_{gs} = \frac{2}{3}C_{ox}WL \tag{10.18}$$

for a transistor in the active region, and that

$$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D} \tag{10.19}$$

We can substitute (10.18) and (10.19) into (10.17), and, after simple manipulation, arrive at

$$\tau_{total} \cong \frac{2nA_0C_{gs}}{g_m} \cong \frac{4}{3}\frac{nA_0L^2}{\mu_n V_{eff}} \tag{10.20}$$

where, again from Chapter 1,

$$V_{eff} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{10.21}$$

Equation (10.20) is very useful in quickly accessing the speed and resolution capabilities of a given technology. It also gives some insight into designing comparators. First, the effective gate-source voltages of input drivers of each stage should be as large as possible to maximize speed. Second, the widths of the transistors have little effect on comparator speed, assuming they are large enough so that C_{gs} dominates parasitic capacitances due to interconnect and external loading. However, larger input transistor sizes will be better matched, resulting in less offset, and will generally have higher transconductance, resulting in lower input-referred noise.

EXAMPLE 10.5

Assume for a 0.2-μm technology that $A_0 = 20$, $n = 3$, $V_{eff} = 0.5$ V, and $\mu_n = 0.03 m^2/V \cdot s$. What is the maximum clocking frequency of the comparator?

Solution

Using (10.20), we have $\tau_{total} = 0.21$ ns. If we assume that 3τ is required for each half period, then a complete period requires 6τ, which corresponds to a clocking frequency of 781 MHz. The resolution is on the order of $2 V/A_0^n = 0.25$ mV.

10.4 LATCHED COMPARATORS

Modern high-speed comparators typically have one or two stages of preamplification followed by a *track-and-latch stage*, as Fig. 10.14 [Yukawa, 1985] shows in simplified form for a CMOS implementation. The rationale

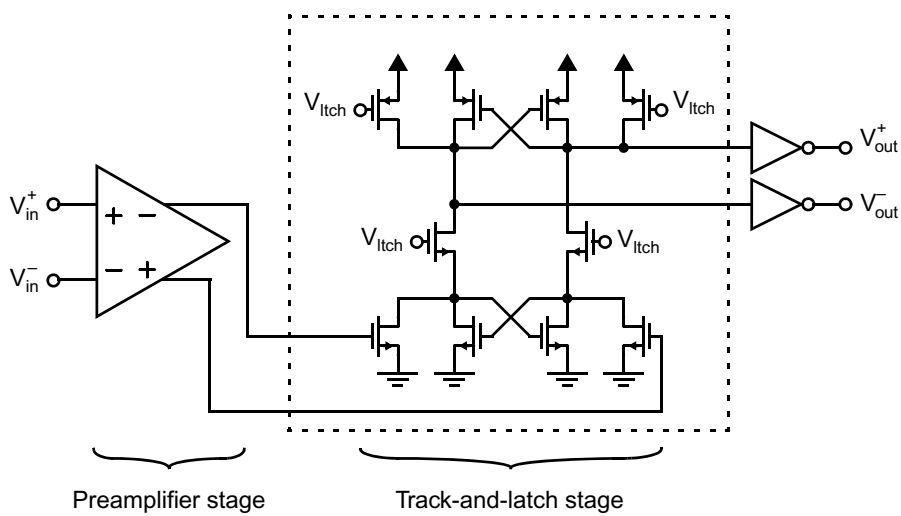


Fig. 10.14 A typical architecture for a high-speed comparator.

behind this architecture is as follows: The preamplifier or preamplifiers are used to obtain higher resolution and to minimize the effects of kickback (which will be explained shortly). The output of the preamplifier, although larger than the comparator input, is still much smaller than the voltage levels needed to drive digital circuitry. The track-and-latch stage then amplifies this signal further during the track phase, and then amplifies it again during the latch-phase, when positive feedback is enabled. The positive feedback regenerates the analog signal into a full-scale digital signal. The track-and-latch stage minimizes the total number of gain stages required, even when good resolution is needed, and, thus, is faster than the multi-stage approach just described.

The preamplifier typically has some gain, perhaps 2 to 10. The preamplifier usually does not have gains much greater than 10; otherwise its time constant is too large and its speed is limited. One or two simple resistively-loaded differential pair stages are often suitable. Note that the input offset and noise of the track-and-latch stage are attenuated by the gain of the preamplifier when referred to the comparator input. Hence, assuming it provides at least a modest gain, the preamplifier's noise and offset will often limit the noise and offset of the comparator.[5] If very high speed but only moderate resolution is required, the preamplifier may sometimes simply be a unity-gain buffer.

It is not good practice to eliminate the preamplifier altogether since kickback into the driving circuitry will then result in very limited accuracy. *Kickback* denotes the charge transfer either into or out of the inputs when the track-and-latch stage goes from track mode to latch mode. This charge transfer is caused by the charge needed to turn the transistors in the positive-feedback circuitry on and by the charge that must be removed to turn transistors in the tracking circuitry off. Without a preamplifier or buffer, this kickback will enter the driving circuitry and cause very large glitches, especially in the case when the impedances seen looking into the two inputs are not perfectly matched.

> ***Key Point:*** *Latched comparators typically employ a simple preamplifier to mitigate kickback and reduce the requirements placed upon the latch stage, followed by a latch which alternates between a reset phase and a positive feedback phase that quickly generates full-swing digital signals from the preamplifier output.*

In high-resolution applications, capacitive coupling and reset switches are also typically included to eliminate any input-offset-voltage and clock-feedthrough errors, in a manner similar to that described in Sections 10.2 and 10.3.

One very important consideration for comparators is to ensure that no memory is transferred from one decision cycle to the next. Recall that when a comparator toggles in one direction, it might have a tendency to stay in that direction, referred to as *hysteresis*. To eliminate hysteresis, one can reset the different stages before entering track mode. This might be achieved by connecting internal nodes to one of the power supplies or by connecting differential nodes together using switches before entering track mode. For example, the comparator shown in Fig. 10.14 has internal nodes of the latch reset to both V_{DD} and ground when the V_{ltch} signal is low. Not only does this eliminate memory, but it also sets the comparator to its *trip point,* which speeds up operation when the comparator resolves small input signals.

Another consideration for comparators is that the gain in track mode should not be too large; otherwise the time constants are too large, which limits the speed.

The track-and-latch stage has many variations. The circuitry shown in Fig. 10.14 should be considered only symbolic, although it will suffice in many applications.

10.4.1 Latch-Mode Time Constant

The time constant of the latch, when it is in its positive-feedback (i.e., latch) phase, can be found by analyzing a simplified circuit consisting of two back-to-back inverters, such as those shown in Fig. 10.15. If we assume that the output voltages of the inverters are close to each other at the beginning of the latch phase, and that the inverters are in their linear range, then each of the inverters can be modelled as a voltage controlled current source driving an

5. The reader may refer to Chapters 2 and 9 to see what determines the random offset and noise, respectively, of simply designed preamplifiers, such as a resistively loaded differential pair.

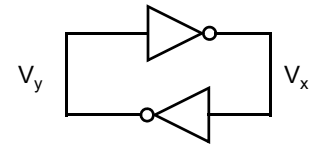


Fig. 10.15 Two back-to-back inverters used as a simplified model of a track-and-latch stage in its latch phase.

RC load, as shown in Fig. 10.16, where A_V is the low-frequency gain of each inverter, which has a transconductance given by $G_m = A_V/R_L$. For this linearized model, we have

$$\frac{A_V}{R_L}V_y = -C_L\left(\frac{dV_x}{dt}\right) - \left(\frac{V_x}{R_L}\right) \tag{10.22}$$

and

$$\frac{A_V}{R_L}V_x = -C_L\left(\frac{dV_y}{dt}\right) - \left(\frac{V_y}{R_L}\right) \tag{10.23}$$

Multiplying (10.22) and (10.23) by R_L and rearranging gives

$$\tau\left(\frac{dV_x}{dt}\right) + V_x = -A_V V_y \tag{10.24}$$

and

$$\tau\left(\frac{dV_y}{dt}\right) + V_y = -A_V V_x \tag{10.25}$$

where $\tau = R_L C_L$ is the time constant at the output node of each inverter. Subtracting (10.25) from (10.24) and rearranging terms gives

$$\left(\frac{\tau}{A_V - 1}\right)\left(\frac{d\Delta V}{dt}\right) = \Delta V \tag{10.26}$$

where $\Delta V = V_x - V_y$ is the voltage difference between the output voltages of the inverters. Equation (10.26) is a first-order differential equation with no forcing function. Its solution is given by

$$\Delta V = \Delta V_0 e^{(A_V - 1)t/\tau} \tag{10.27}$$

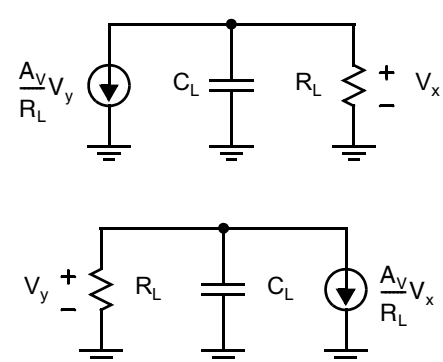


Fig. 10.16 A linearized model of the track-and-latch stage when it is in its latch phase.

where ΔV_0 is the initial voltage difference at the beginning of the latch phase. Thus, the voltage difference increases exponentially in time with a time constant given by

$$\tau_{ltch} = \frac{\tau}{A_V - 1} \cong \frac{R_L C_L}{A_V} = \frac{C_L}{G_m} \tag{10.28}$$

where, again, $G_m = A_V / R_L$ is the transconductance of each inverter. Note that τ_{ltch} is roughly equal to the inverse of the unity-gain frequency of each inverter.

In the case of MOS devices, normally the output load is proportional to the gate-source capacitance of a single transistor, or specifically,

$$C_L = K_1 W L C_{ox} \tag{10.29}$$

Here, K_1 is a proportionality constant, usually between 1 and 2. Also, the inverter transconductance is proportional to the transconductance of a single transistor, and is given by

$$G_m = K_2 g_m = K_2 \mu_n C_{ox} \frac{W}{L} V_{eff} \tag{10.30}$$

where K_2 might be between 0.5 and 1. Substituting (10.29) and (10.30) into (10.28) gives

$$\tau_{ltch} = \frac{K_1}{K_2 \mu_n V_{eff}} \frac{L^2}{} = K_3 \frac{L^2}{\mu_n V_{eff}} \tag{10.31}$$

where K_3 might be between 2 and 4. Note that (10.31) implies that τ_{ltch} depends primarily on the technology and not on the design (assuming a reasonable choice of V_{eff} and that C_L is minimized). Note also the similarity between (10.31) and (10.20), the equation for the time constant of a cascade of gain stages. For a given technology, (10.31) is very useful in determining a rough estimate for the maximum clock frequency of a latch-and-track comparator.

If it is necessary for a voltage difference of ΔV_{logic} to be obtained in order for succeeding logic circuitry to safely recognize the correct output value, then, by using (10.27), we find the time necessary for this to happen is given by

$$T_{ltch} = \frac{C_L}{G_m} \ln\left(\frac{\Delta V_{logic}}{\Delta V_0}\right) = K_3 \frac{L^2}{\mu_n V_{eff}} \ln\left(\frac{\Delta V_{logic}}{\Delta V_0}\right) \tag{10.32}$$

If ΔV_0 is small, this latch time can be large, perhaps larger than the allowed time for the latch phase. Such an occurrence is often referred to as *metastability*. In other words, because its initial value is too small, the differential output voltage of the latch does not increase enough to be recognized as the correct logic value by succeeding circuitry. Sometimes, even when the initial voltage difference is large enough, it is possible that circuit noise can cause the initial voltage difference to become small enough to cause metastability.

Key Point: *A positive-feedback latch output grows exponentially with a time constant given by the unity-gain frequency, G_m/C_L. When the initial input is too small, metastability may occur whereby the growth is too slow for a recognizable logic signal to appear at the latch output before the onset of the reset phase.*

EXAMPLE 10.6

Assume for a 0.2-μm technology that $K_3 = 2.5$, $\Delta V_0 = 1$ mV, $\Delta V_{logic} = 1$ V, $V_{eff} = 0.5$ V, and $\mu_n = 0.03\ M^2/V \cdot s$. What is the maximum clocking frequency of the comparator?

Solution

Using (10.32), we have T_{ltch} = 46ps. Assuming this is half the total period (which might be quite an optimistic assumption), the smallest possible period for a track-and-latch might be 92 ps, which corresponds to a clocking frequency of 11 GHz. In a typical comparator, the maximum clock rate would probably be limited by the frequency response of the preamplifiers and the track-and-latch during the track phase, rather than by the speed of the latch during the latch phase.

10.4.2 Latch Offset

The input offset of a latch is influenced by random and systematic mismatch between its two half-circuits. When very small transistors are used, this mismatch may be significant often leading of offset of 10's of mV or more. In such cases, additional components may be introduced to so that by appropriately controlling the imbalances in the circuit, the effects of random mismatch can be cancelled, resulting in a latch with near-zero overall offset.

The latch model in Fig. 10.16 is modified in Fig. 10.17 to illustrate several contributors to latch offset. The imbalances $i_{o1} \neq i_{o2}$, $G_{m1} \neq G_{m2}$, etc. may be due to random mismatch and/or they may be introduced intentionally to provide a means of adjusting the latch offset. Clearly, when $i_{o1} \neq i_{o2}$ and/or $G_{m1} \neq G_{m2}$ some offset is introduced since, with $V_x = V_y \neq 0$, different currents flow in each half circuit. However, even if the offset current i_o and transconductances G_m are perfectly matched and $V_x(0) = V_y(0) \neq 0$, if the R_L loads are mismatched the node voltages V_x and V_y will respond with different time constants immediately giving rise to a finite differential voltage $\Delta V = V_x = V_y$ which in turn initiates the circuit's positive feedback, eventually regenerating a logic signal. The evolution of the node voltages in this circumstance is illustrated in Fig. 10.18.

Key Point: *Latches have offset due to imbalances between the two half-circuits providing positive feedback. Imbalances arise due to random mismatch. Additional controllable imbalances are sometimes included by design in order to cancel random offsets and provide an accurate comparator.*

A latch incorporating an offset adjustment mechanism is shown in Fig. 10.19 [Cho, 1995]. In latch mode, the cross-coupled pair $Q_{1,2}$ is source-degenerated by the triode transistors, $Q_{3\text{-}6}$. Hence the transconductance of each half-circuit, $G_{m1,2}$, depends upon the applied differential voltage $V_{ref}^{\pm}$ and the input voltage $V_{in}^{\pm}$. A dc current i_o flows at the onset of latch mode since the node voltages $V_{out}^{\pm}$ have been reset to the positive supply, and all NMOS transistors are turned on. Hence, in the absence of any random mismatch, the latch will be in a balanced state with equal half-circuit currents flowing only when

$$\Delta V_{in} = \left(\frac{W_{3,6}}{W_{4,5}}\right)\Delta V_{ref} \tag{10.33}$$

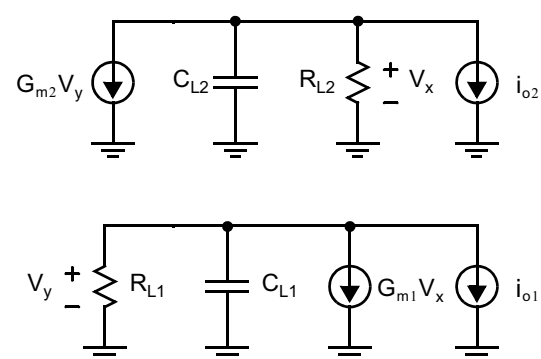


Fig. 10.17 A linearized model of the latch illustrating various sources of offset, random and/or intended.

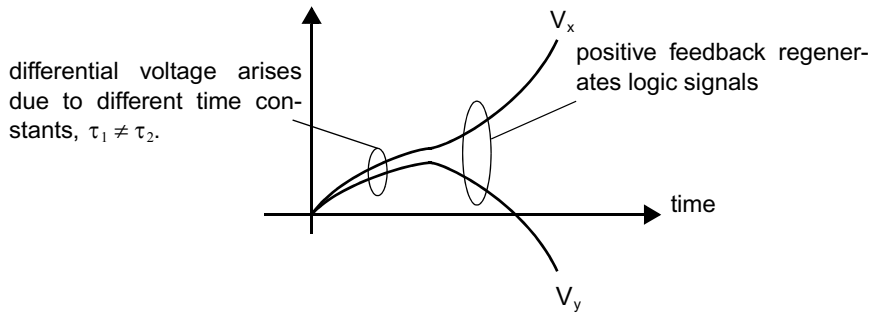


Fig. 10.18 Evolution of the latch node voltages when $i_{o1} = i_{o2} \neq 0$ and $R_{L1}C_{L1} \neq R_{L2}C_{L2}$.

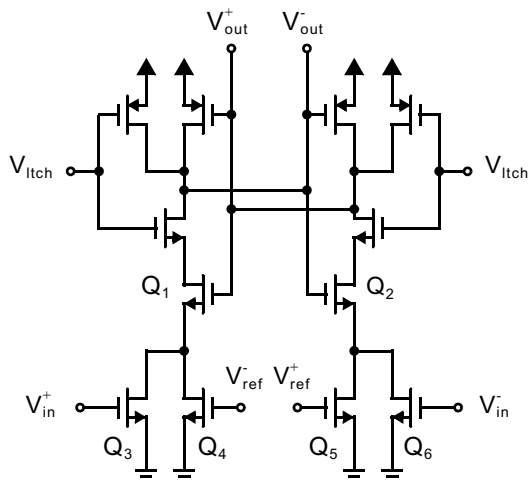


Fig. 10.19 A latch circuit with controllable offset [Cho, 1995].

Equation (10.33) is the latch offset due ΔV_{ref}, which may be adjusted to cancel other sources of offset, such as those arising from mismatches in device sizes, parasitics, etc. A rigorous general analysis relating all sources of offset to the circuit parameters is quite complex [He, 2009].

10.5 EXAMPLES OF CMOS AND BiCMOS COMPARATORS

This section describes a number of possible high-speed comparators.

The literature on integrated circuit technology has many examples of latched comparators. For example, either of the latches shown in Fig. 10.14 and Fig. 10.19 may be preceded by a resistively loaded differential pair preamplifier to realize a latched comparator. Figure 10.20 shows another comparator [Song, 1990]. It has the positive feedback of the second stage always enabled. In track mode, when the two diode-connected transistors of the gain stage are enabled, the gain around the positive-feedback loop is less than one, and the circuit is stable. The combination of the diode-connected transistors of the gain stage and the transistors of the positive-feedback loop acts as a moderately large impedance and gives gain from the preamplifier stage to the track-and-latch stage. The diode-connected loads of the preamplifier stage give a limited amount of gain in order to maximize speed, while still buffering the kickback from the input circuitry.

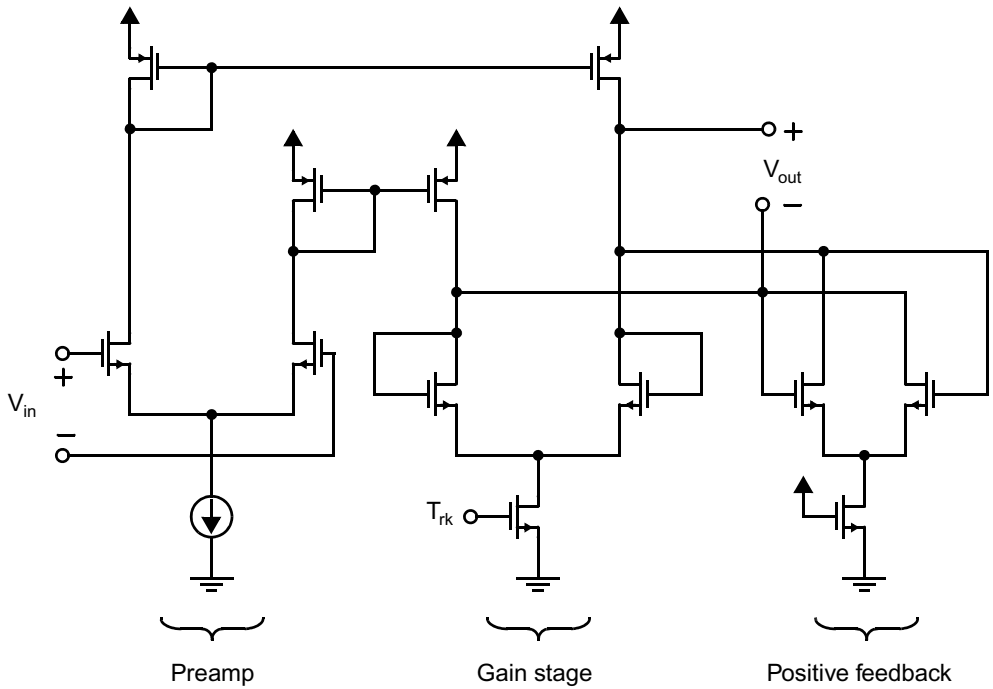


Fig. 10.20 A two-stage comparator that has a preamplifier and a positive-feedback track-and-latch stage.

Another comparator is shown in Fig. 10.21 [Norsworthy, 1989]. This design also uses diode-connected loads to keep all nodes at relatively low impedance (similar to current-mode circuit-design techniques), thereby keeping all node time constants small and giving fast operation. It uses precharging to eliminate any memory from the previous decision that might lead to hysteresis. For example, the positive-feedback stage is precharged low, whereas the digital-restoration stage is precharged high, somewhat similar to what is done in Domino CMOS logic [Krambeck, 1982].

Yet another comparator is shown in Fig. 10.22 with appropriate clock waveforms.[6] This design eliminates any input-offset voltages from both the first and second stages by using capacitive coupling. It also has common-mode feedback circuitry for the first preamplifier stage, which allows for input signals that have large common-mode signals. Unlike in fully differential opamps, in fully differential comparators, the linearity of the common-mode feedback circuitry is noncritical since, whenever large signals are present (and the common-mode feedback circuitry becomes nonlinear), there is no ambiguity in resolving the sign of the input signal. This allows a simple differential pair to be used for CMFB circuitry. Measured (but unpublished) performance of this circuit resulted in a 0.1-mV resolution at a 2-MHz clock frequency, despite a very old 5-μm technology. The performance measured was limited by the test setup rather than by the circuitry.

A BiCMOS comparator that exhibits very good performance is described in [Razavi, 1992]. It is based on the idea that it is not necessary to reset the first stage; the input-offset errors of the first stage can still be eliminated by resetting the right side of the coupling capacitors between the first stage and the second stage, assuming that the gain of the first stage is not too large [Poujois, 1978] [Vittoz, 1985]. A simplified schematic of this comparator is shown in Fig. 10.24. During the reset phase, the inputs to the preamplifier are connected directly to

6. K. Martin, 1984 (experimentally verified, but unpublished).

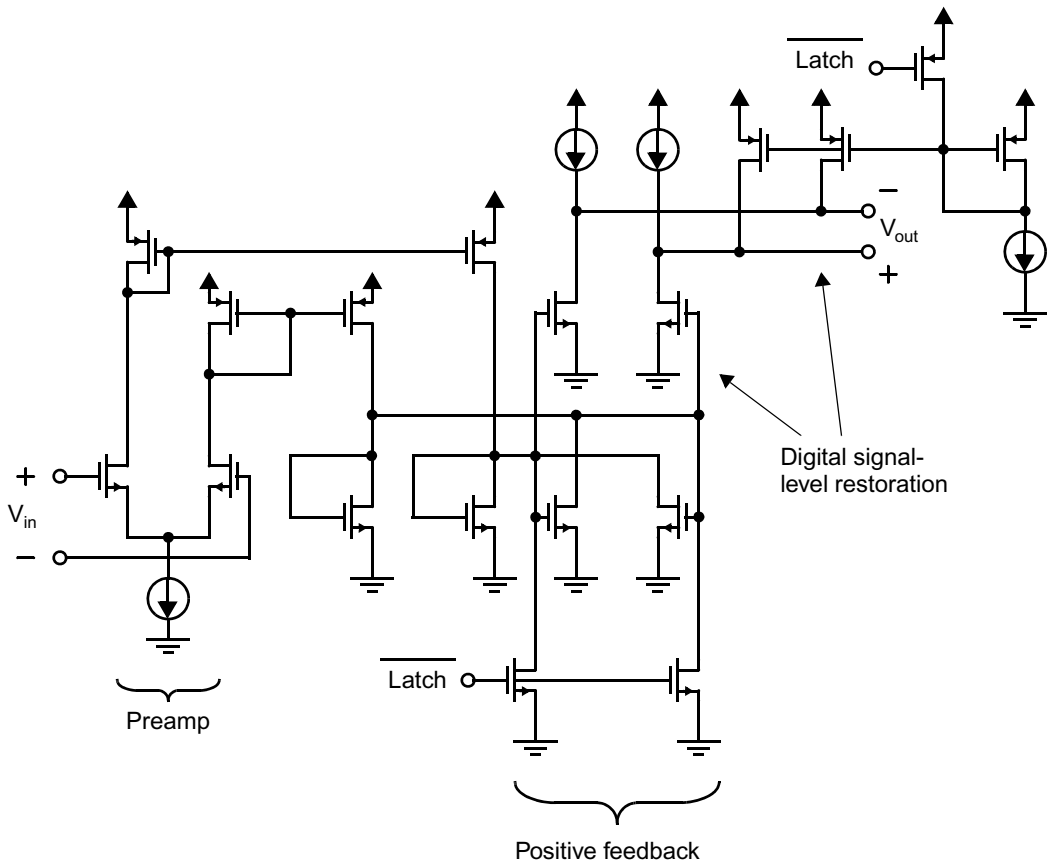


Fig. 10.21 A two-stage comparator.

ground (or to a reference voltage), and the outputs of the coupling capacitors are connected to ground as well. This stores any offset voltages of the first stage on the capacitors. When the comparator is taken out of reset phase, then the effect of the clock feedthrough of S_5 and S_6 on the input resolution is divided by the gain of the first stage. In the realization described in [Razavi, 1992], the first stage is a BiCMOS preamplifier, consisting of MOS source followers followed by a bipolar differential amplifier and emitter-follower output buffers, as shown in Fig. 10.25. Notice that the circuit operates between ground and a negative voltage supply. Note also that the switches are realized using p-channel transistors. The track-and-latch stage has both a bipolar latch and a CMOS latch, as shown in Fig. 10.23. During the reset phase, ϕ_1 is low and ϕ_2 is high, X_1 and Y_1 are connected to ground through switches M_1 and M_2, C_3 is discharged to the minus supply through M_{11}, and X_2 and Y_2 are discharged to the minus supply through M_9 and M_{10}. Also, at this time, M_3 and M_4 are off. Next, ϕ_1 becomes high, which leaves X_1 and Y_1 floating and connects the outputs of the preamplifier to its input signal. After a short delay, which is needed for the transient response of the preamplifier, ϕ_2 becomes low, which turns on M_{12}, thereby activating the positive feedback action of Q_5 and Q_6. This develops a differential voltage between X_1 and Y_1 of about 200 mV, since C_3 is about one-fifth the size of C_1 and C_2. The offset of this bipolar latch is quite small, typically on the order of one millivolt or less. Also note that, when activated, the only power dissipated is that required to charge C_3—there is no dc power dissipation. A short time after the bipolar latch is activated, the inverter connected to C_3 changes state, and its output voltage will drop. This low voltage turns on M_3 and M_4, which activates the CMOS part of the latch consisting of cross-coupled M_5 and M_6 and cross-coupled

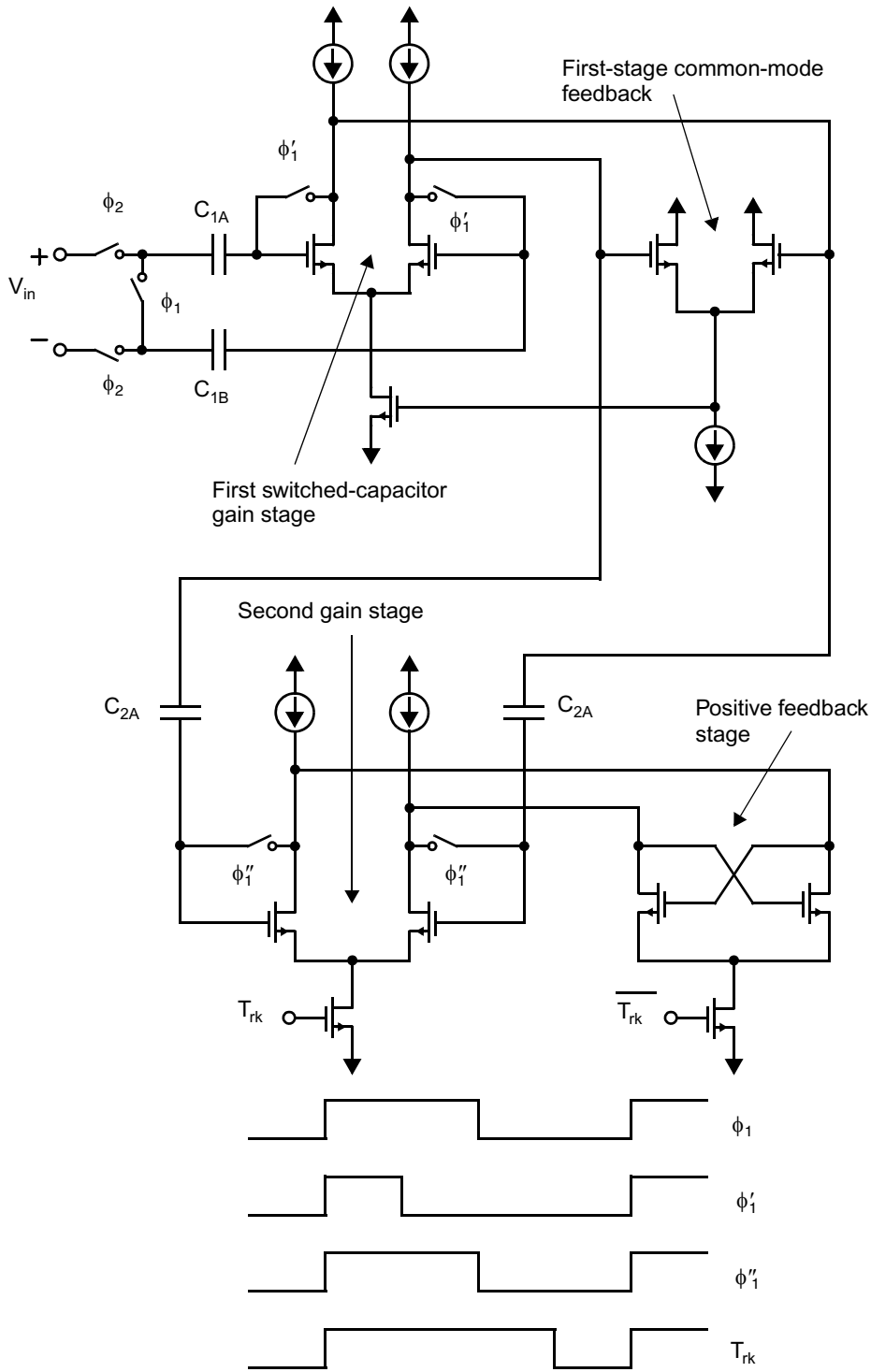


Fig. 10.22 A two-stage comparator with capacitive coupling to eliminate input-offset-voltage and clock-feedthrough errors, along with positive feedback for fast operation.

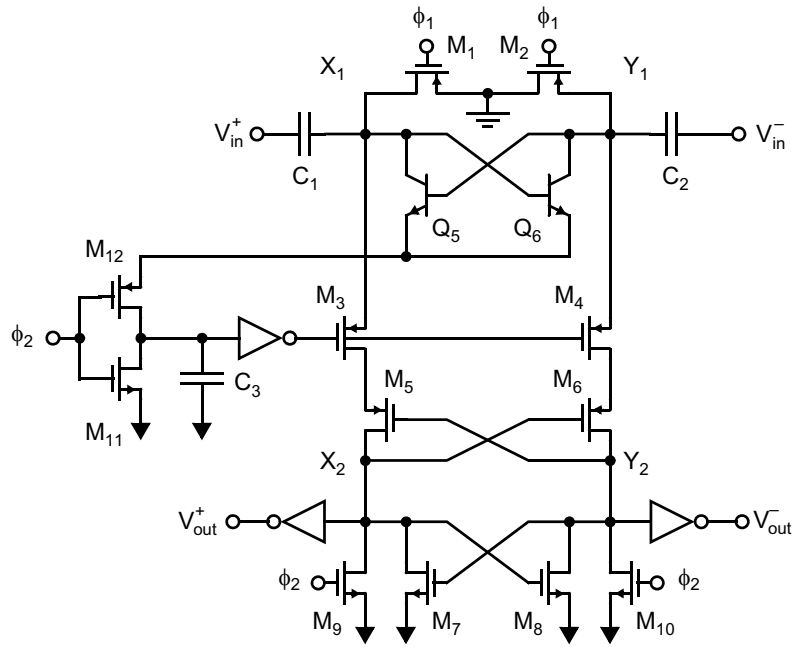


Fig. 10.23 The BiCMOS track-and-latch stage in [Razavi, 1992].

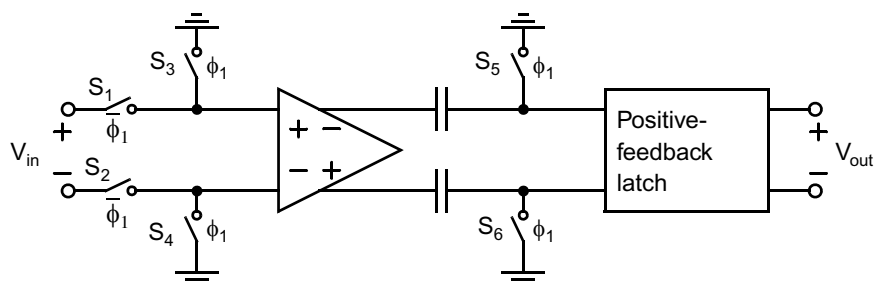


Fig. 10.24 A simplified schematic of the comparator described in [Razavi, 1992].

M_7 and M_8. The reasons for including cross-coupled M_5 and M_6 are to prevent nodes X_1 and Y_1 from being discharged very far from ground toward the minus power supply and at the same time to amplify the 200-mV difference signal developed by the bipolar latch to almost a full-level CMOS voltage change. Thus, the latch is accurate, relatively fast, and low power. Also described in [Razavi, 1992] is an interesting CMOS-only comparator, which is not described here, but which interested readers may investigate.

10.5.1 Input-Transistor Charge Trapping

When realizing very accurate CMOS comparators, an error mechanism that must be considered is charge trapping in the gate oxide of the input transistors of the MOS comparators [Tewksbury, 1989]. When n-channel transistors are stressed with large positive gate voltages, electrons can become trapped via a tunneling mechanism in which

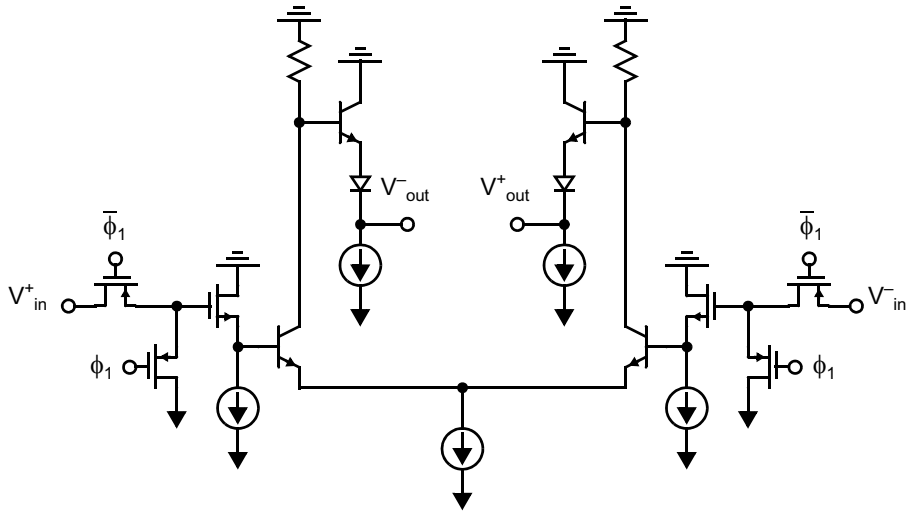


Fig. 10.25 A BiCMOS preamplifier.

electrons tunnel to oxide traps close to the conduction band. The time constant for the release of these trapped electrons is on the order of milliseconds, much longer than the time it takes for them to become trapped. During the time they are trapped, the effective transistor threshold voltage is increased. This leads to a comparator hysteresis on the order of 0.1 to 1 mV. This effect correlates well with transistor 1/f noise and is much smaller in p-channel transistors. When very accurate comparators are needed, charge trapping must be considered and minimized. One possible means of minimizing this effect, assuming we have a BiCMOS technology, is described in [Miller, 1990] and is illustrated in Fig. 10.26. A BiCMOS technology allows diodes to be used in the input

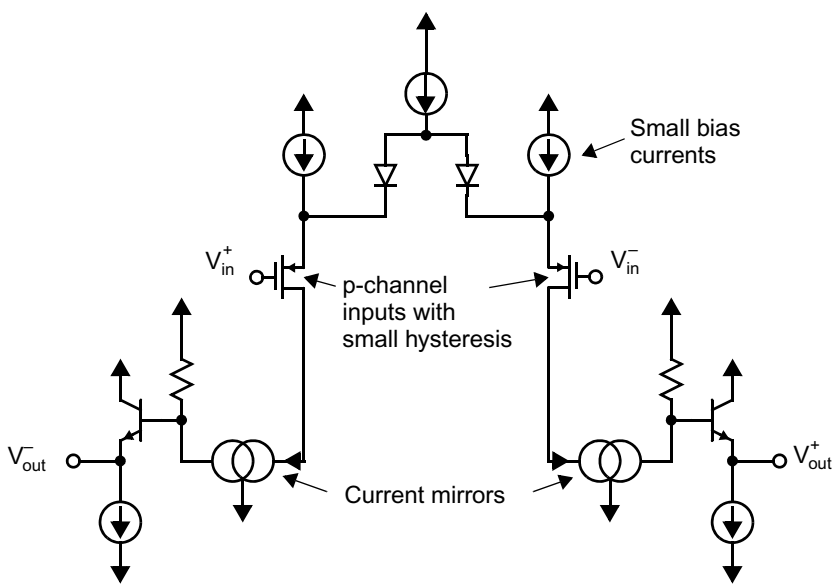


Fig. 10.26 A comparator with little hysteresis. (Hysteresis is caused by charge trapping.)

stage, which allows the inclusion of two additional small-current sources in the input stage; this guarantees that the p-channel input transistors never turn off.[7] Also, by using p-channel input transistors, the charge trapping is greatly minimized because p-channel transistors exhibit much less hysteresis than n-channel transistors.

An alternative approach is to flush the input transistors after each use where the junctions and wells of n-channel transistors are connected to a positive power supply whereas the gates are connected to a negative power supply. This effectively eliminates the trapped electrons [Swanson, 1993]. Alternatively, or in addition, two input stages can be used for the comparator—a rough stage can be used during times when large signals with overloads are possible, whereas a fine stage can be used during times when accurate comparisons are necessary and it can be guaranteed that no large signals are present [Swanson, 1993; Tan, 1990].

10.6 EXAMPLES OF BIPOLAR COMPARATORS

High-speed bipolar comparators are typically latched comparators, as shown previously in Fig. 10.14, where preamplifier and track-and-latch stages are used. A typical example is shown in Fig. 10.27 [Wakimoto, 1988]. The preamplifier is a simple differential amplifier that might have a gain of around 5. Besides increasing the resolution, this preamplifier helps to eliminate kickback from the track-and-latch stage. The track-and-latch stage is very similar to a current-mode digital latch. During track phase, the differential pair consisting of transistors Q_3 and Q_4 is enabled and operates as a second differential amplifier. When latch mode is enabled, I_2 is diverted to the differential pair consisting of Q_5 and Q_6. The inputs to this third differential pair come

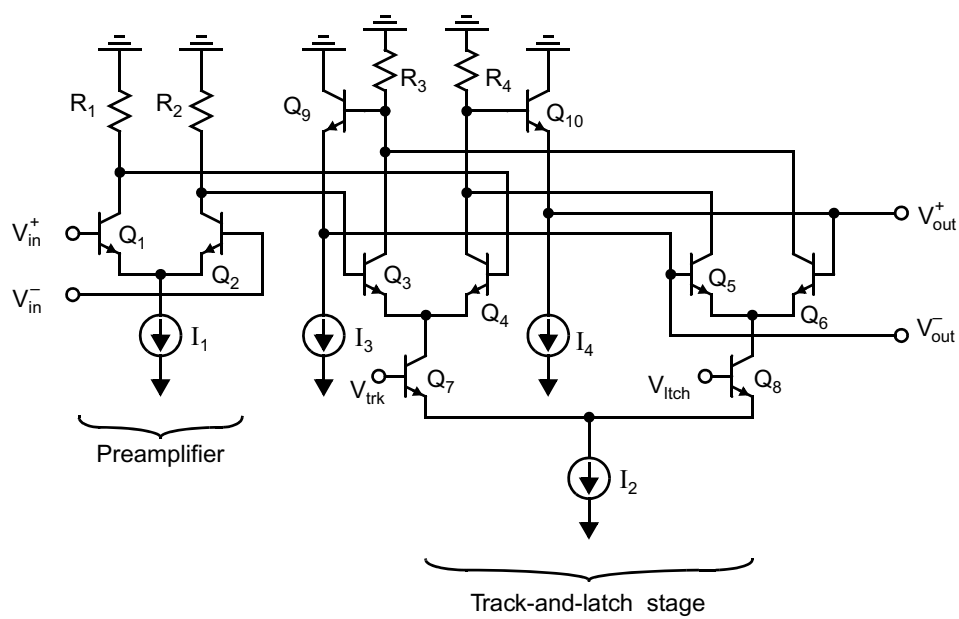


Fig. 10.27 A typical bipolar comparator.

7. p-channel transistors exhibit trapping for positive gate voltages since only electrons, and not holes, are trapped. Therefore, to prevent trapping, they should never be turned off.

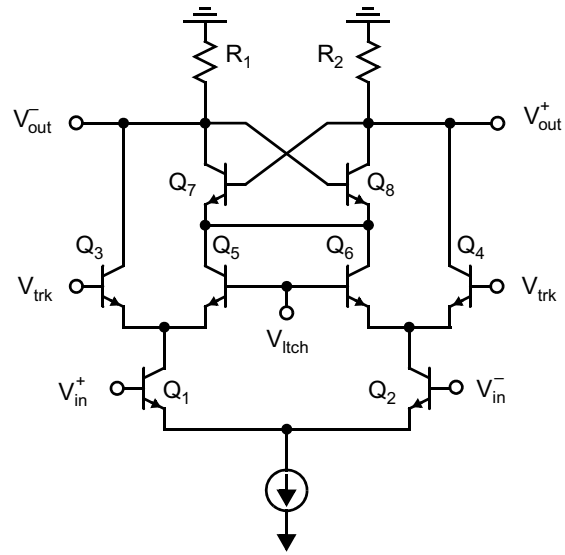


Fig. 10.28 A second bipolar comparator.

from the outputs of the second differential pair through emitter-follower buffers consisting of Q_9 and Q_{10}. The outputs of the third differential pair are connected back to the high-impedance output nodes in a cross-coupled manner so that positive feedback results. When this positive feedback is enabled, it takes the initially-small differential voltage and quickly amplifies it to between 0.4 and 0.6 V (depending on the logic levels being used). There are two reasons for using the emitter-follower buffers. These buffers isolate the high-impedance nodes at the collectors of the second and third differential pairs from the base-emitter capacitances of the third differential pair. They also isolate the same high impedance nodes from the output load capacitance.

Another architecture for a bipolar comparator is shown in Fig. 10.28 [Van de Plassche, 1988]. In this comparator, the input signals go to a differential pair that is always turned on (even during the latch phase). During the track phase, Q_3 and Q_4 are on. They operate as common-base transistors, and therefore the comparator operates as a cascode amplifier during this phase. When the comparator goes into latch mode, the collector current from the input transistors is redirected from the common-base transistors to transistors Q_7 and Q_8, which are connected in a positive-feedback loop. Since the bias current of the input transistors is never turned off, the kickback from them is greatly reduced without adding additional preamplifier stages. This maximizes the bandwidth during the tracking phase.

A third bipolar comparator is shown in Fig. 10.29. Although this comparator has never been experimentally tested, it is used as an example to illustrate many comparator design principles. A small current source is added to the tracking differential pair so that this differential pair doesn't turn off completely when the comparator is in hold mode. This makes the comparator faster when it goes from hold mode into track mode. Also, a cascode common-base stage is added just below the output voltages. This stage is used to isolate the outputs from the positive-feedback loop. This in turn allows for much smaller resistors to be used in the positive-feedback loop, which speeds this loop up and eliminates the need for emitter-follower buffers in the loop. Finally, a larger resistor is added between the emitters of the cascode transistors. This resistor prevents the cascode transistors from turning off completely during the latch phase, which also speeds up the transition from latch mode to tracking mode.

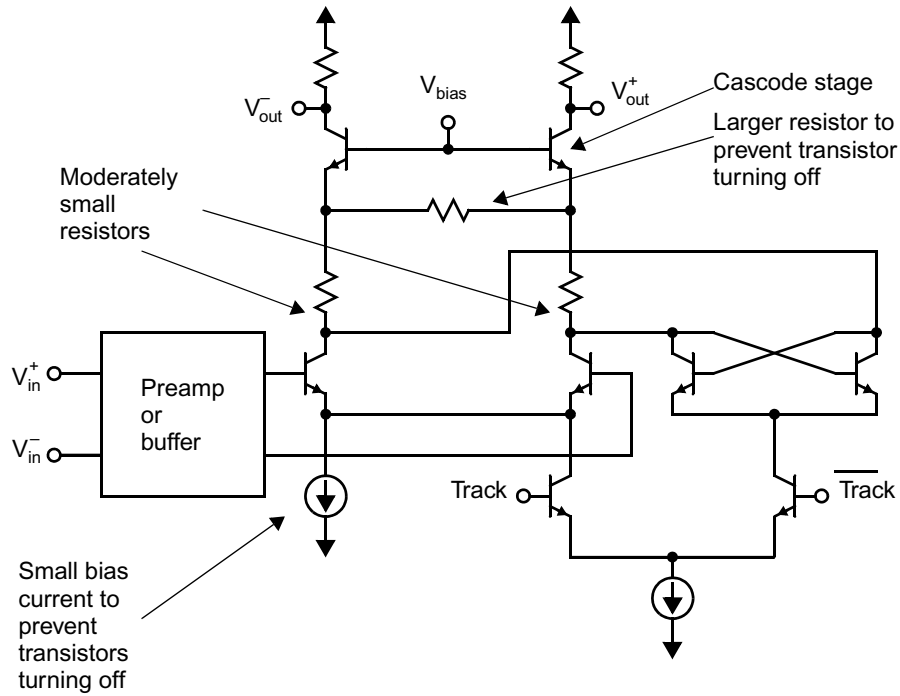


Fig. 10.29 A third bipolar comparator.

10.7 KEY POINTS

- The input offset of a comparator is the input at which the output is equally likely to be high or low. Noise determines how the output statistics vary as a function of input. [p. 413]
- Hysteresis is "memory" in a comparator that tends to make it remain in its past output state, usually due to charge retained on internal circuit capacitances. [p. 414]
- Regular opamps are compensated to ensure stability in closed-loop configurations, resulting in a low 3-dB bandwidth and, hence, very slow settling if used open-loop as a comparator. [p. 417]
- An input capacitor is sometimes used to sample the input offset voltage of a comparator during a reset phase, and then subtract it from the input during the comparison phase. [p. 417]
- As a MOSFET goes from triode mode to off, the charge stored in the channel region must flow out of the transistor terminals; in a comparator, the charge may accumulate on circuit capacitances causing offset. [p. 420]
- The deleterious effects of charge-injection can be reduced by using "advanced" clocks which open-circuit the discharge paths through which signal-dependent charge-injection flows. [p. 421]
- Latched comparators typically employ a simple preamplifier to mitigate kickback and reduce the requirements placed upon the latch stage, followed by a latch which alternates between a reset phase and a positive feedback phase that quickly generates full-swing digital signals from the preamplifier output. [p. 427]
- A positive-feedback latch output grows exponentially with a time constant given by the unity-gain frequency, G_m/C_L. When the initial input is too small, metastability may occur whereby the growth is too slow for a recognizable logic signal to appear at the latch output before the onset of the reset phase. [p. 429]
- Latches have offset due to imbalances between the two half-circuits providing positive feedback. Imbalances arise due to random mismatch. Additional controllable imbalances are sometimes included by design in order to cancel random offsets and provide an accurate comparator. [p. 430]

10.8 REFERENCES

T. B. Cho and P. R. Gray. "A 10b, 20 Msample/s, 35 mW Pipeline A/D Converter," *IEEE J. Solid-State Circuits*, Vol. 30, pp. 166–172, Mar. 1995.

D. G. Haigh and B. Singh. "A Switching Scheme for Switched-Capacitor Filters, Which Reduces Effect of Parasitic Capacitances Associated with Control Terminals," *Proc. IEEE Int. Symp. on Circuits and Systems,* Vol. 2, pp. 586–589, June 1983.

H. He, S. Zhan, D. Chen, R. L. Geiger. "Analyses of Static and Dynamic Random Offset Voltages in Dynamic Comparators," *IEEE Trans. Circuits and Systems I: Regular Papers*, Vol. 56, pp. 911–919, May 2009.

R. H. Krambeck. C. Lee, and H. S. Law. "High-Speed Compact Circuits with CMOS," *IEEE J. of Solid-State Circuits,* Vol. 17, no. 3, pp. 614–619, June 1982.

K. Martin. "Improved Circuits for the Realization of Switched-Capacitor Filters," *IEEE Trans. Circuits and Systems*, Vol. CAS-27, no. 4, pp. 237–244, April 1980.

———. "New Clock-Feedthrough Cancellation Technique for Analog MOS Switched Capacitor," *Electron. Lett*., no. 18, 1992.

K. Martin, L. Ozcolak, Y. S. Lee, and G. C. Temes. "A Differential Switched-Capacitor Amplifier," *IEEE J. of Solid-State Circuits*, Vol. SC-22, no. 1, pp. 104–106, February 1987.

L. L. McCreary and P. R. Gray. "All-MOS Charge Redistribution Analog-to-Digital Conversion Techniques—Part I," *IEEE J. of Solid-State Circuits*, Vol. SC-10, pp. 371–379, December 1975.

G. Miller, M. Timko, H. S. Lee, E. Nestler, M. Mueck, and P. Ferguson. "An 18-b 10-μs Self-Calibrating ADC," *IEEE Intern. Solid-State Circuits Conf.*, pp. 168–169, February 1990.

S. Norsworthy, I. Post, and S. Fetterman. "A 14-bit 80-KHz Sigma-Delta A/D Converter: Modeling, Design, and Performance Evaluation," *IEEE J. of Solid-State Circuits*, Vol. 24, no. 2, pp. 256–267, April 1989.

R. Poujois and J. Borel. "A Low-Drift Fully Integrated MOSFET Operational Amplifier," *IEEE J. of Solid-State Circuits*, Vol. SC-13, pp. 499–503, August 1978.

B. Razavi. *Principles of Data Conversion System Design*, IEEE Press, Piscataway, New Jersey, 1995.

B. Razavi and B. Wooley. "Design Techniques for High-Speed, High-Resolution Comparators," *IEEE J. of Solid-State Circuits*, Vol. 27, no. 12, pp. 1916–1926, December 1992.

J. Shieh, M. Patil, and B. L. Sheu. "Measurement and Analysis of charge-injection in MOS Analog Switches," *IEEE J. of Solid-State Circuits*, Vol. 22, no. 2, pp. 277–281, April 1987.

B. Song, H. S. Lee, and M. Tompsett. "A 10-b 15-MHz CMOS Recycling Two-Step A/D Converter," *IEEE J. of Solid-State Circuits*, Vol. 25, no. 6, pp. 1328–1338, December 1990.

E. J. Swanson. "Method and Circuitry of Decreasing the Recovery Time of a MOS Differential Voltage Comparator," U.S. patent no. 5, 247, 210, September 1993.

K. Tan, et al. "Error-Correction Techniques for High-Performance Differential A/D Converters," *IEEE J. of Solid-State Circuits,* Vol. 25, no. 6, pp. 1318–1327, December 1990.

T. L. Tewksbury, H. S. Lee, and G. Miller. "The Effects of Oxide Traps on the Large-Signal Transient Response of Analog MOS Circuits," *IEEE J. of Solid-State Circuits,* Vol. 24, no. 2, pp. 542–543, April 1989.

R. Van de Plassche and P. Baltus. "An 8-bit 100-MHz Full-Nyquist Analog-to-Digital Converter," *IEEE J. of Solid-State Circuits*, Vol. 23, no. 6, pp. 1334–1344, December 1988.

E. A. Vittoz. "Dynamic Analog Techniques," in *Design of MOS VLSI Circuits for Telecommunications*, ed. Y. Tsividis and P. Antognetti, Prentice Hall, Englewood Cliffs, New Jersey, 1985.

T. Wakimoto, Y. Akazawa, and S. Konaka. "Si Bipolar 2-GHz 6-bit Flash A/D Conversion LSI," *IEEE J. of Solid-State Circuits,* Vol. 23, no. 6, pp. 1345–1350, December 1988.

Y. S. Yee, L. M. Terman, and L. G. Heller. "A 1-mV MOS Comparator," *IEEE J. of Solid-State Circuits*, Vol. SC-13, pp. 294–297, June 1978.

Yubawa, A. "A CMOS 8-bit High-Speed A/D Converter," *IEEE J. of Solid-State Circuits*, Vol. SC-20, no. 3, pp. 775–779, June 1985.

10.9 PROBLEMS

Unless otherwise stated, assume the following:

- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8\ V$

$\gamma = 0.5\ V^{1/2}$
$r_{ds}(\Omega) = 8000\ L\ (\mu m)/I_D\ (mA)$ in active region
$C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
$C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
$C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
$C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$

- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9\ V$
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}(\Omega) = 12{,}000\ L\ (\mu m)/I_D\ (mA)$ in active region
  $C_j = 4.5 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.5 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$

10.9.1 Section 10.1: Comparator Specifications

10.1 A comparator has an input offset of 10 mV and input noise of 0.2 mV_{rms}. What is the smallest input that can be resolved with an error rate of less than 10^{-12}?

10.2 The output of a comparator is observed for various input voltages resulting in the plot of Fig. P10.2. What is the input offset and noise?

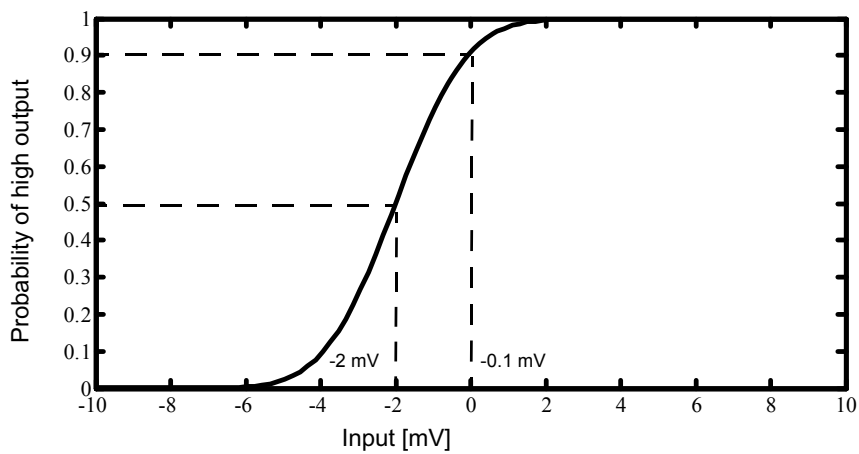


Fig. P10.2

10.9.2 Section 10.2: Using an Opamp for a Comparator and Section 10.3: Charge-Injection Errors

10.3 The CMOS transmission gate shown in Fig. P10.3 has an input voltage of 2.5 V when it turns off. The W/L of the n-channel is 5 μm/0.8 μm and the W/L of the p-channel is 15 μm/0.8 μm. Assume the total parasitic capacitance between the output node and ground is 50 fF, V_{DD} = 5 V, and that the clock signals change very fast. Also ignore changes due to overlap capacitance. Estimate the change in output voltage due to charge-injection. What will the final output voltage be?

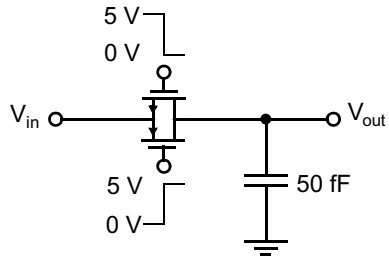


Fig. P10.3

10.4 A simple CMOS inverter is used as a comparator, as shown in Fig. P10.4. Assume ϕ_{1a} drops before ϕ_1 drops, so that only the charge-injection from Q_1 need be considered. What is the change in V_{out} due to this charge-injection? Assume the transistors in the inverter have sizes chosen such that the inverter threshold voltage is at $V_{DD}/2 = 2.5$ V (i.e., $W_5/W_4 = \mu_n/\mu_p$), and that all transistor lengths are 0.8 μm. Also assume the inverter gain is –24 at $V_{out} = 2.5$ V, and ignore overlap capacitance. Do not ignore the body effect on the threshold voltage of Q_1.

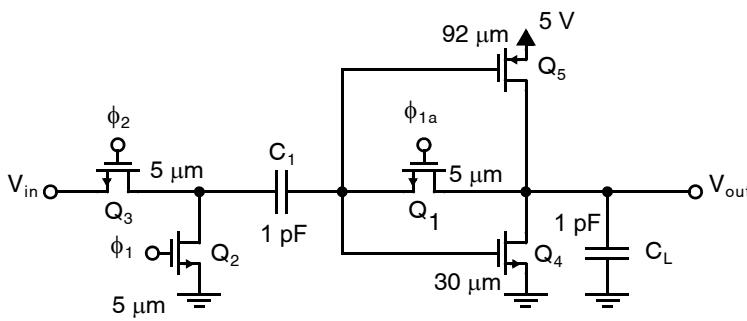


Fig. P10.4

10.5 Assume that three stages of inverters identical to the inverter used in Problem 10.4 are cascaded as shown in Fig. 10.11. What is the input-voltage offset error due to the charge-injection of the reset switch of the last stage only?

10.6 Estimate the time constant for the inverter used in Problem 10.4 during its reset phase and also during its comparison phase.

10.7 For the circuit shown in Fig. 10.11, find the voltages at each of the opamp inputs at the end of ϕ_i', ϕ_i'', ϕ'', ϕ_1, and ϕ_2, where V_i is the input and $V_{err\text{-}i}$ is the error due to the reset switch of the $_i$th opamp. Assume opamps are offset free and that they remain in their linear region after their reset switch is opened. Show that the overall input-referred clock feedthrough is given by $|V_{err\text{-}3}|/(A_1A_2)$.

10.8 Repeat Problem 10.7 where offsets of the opamps are not ignored. What is the input-referred offset and clock feedthrough?

10.9 Assume a MOS transistor's unity-gain frequency is approximately given by $\omega_t = g_m/C_{gs}$. Find an expression for ω_t in terms of V_{eff} and C_{ox}. Assume a single gain stage without compensation has a unity-gain frequency given by $\omega_a = \omega_t/2$. Further assume that a comparator is realized by a cascade of three stages, each having a gain of –15. If three time constants are required for settling during the open-loop comparison phase, what is the maximum frequency of operation for the comparator? What is the resolution? You may assume that $V_{eff} = 0.25$ V, $\mu_n = 0.05$ $M^2/V \cdot s$, and that there are only n-channel transistors in the signal path.

10.9.3 Sections 10.4–10.6: Latched Comparators and CMOS, BiCMOS, and Bipolar Examples

10.10 A positive-feedback latch is found to have an input offset of 5 mV, noise of 0.5 mV, and hysteresis of 2 mV. A preamplifier is to be placed in front of the latch so that the resulting comparator has a overall input offset of no more than 3 mV, input noise of no more than 0.4 mV, and hysteresis of no more than 1 mV. Find specifications on the gain, input offset, and input-referred noise of the pre-amplifier. Design a resistively-loaded MOS differential pair to meet these specifications (under 3σ mismatch conditions). Use the 0.18-μm CMOS process in Table 1.5 and assume the following mismatch parameters: $A_{Vt0} = 4\ \text{mV}\cdot\mu\text{m}$ and $A_{K'} = 0.01\ \mu\text{m}$.

10.11 Give reasonable device widths of the comparator in Fig. 10.20 so that the first stage has a gain of 5 and so that the latch-and-track stage has a gain of 10 when in track mode and a loop gain of 4 when in latch mode. Assume the input transistors are 25-μm wide and that the bias current of the first stage is 0.1 mA.

10.12 Assume the comparator of Fig. 10.20 has the device sizes found in Problem 10.11. Estimate the equivalent time constant when the comparator is in track mode. If the input signal changes from 0 to 5 mV at time 0, how long would it take the differential output voltage to change by 50 mV?

10.13 Consider the comparator in Fig. 10.20 with the device sizes found in Problem 10.11. Estimate the equivalent time constant of the latch-and-track stage when the comparator is in latch mode. If the differential output voltage is 50 mV at the beginning of the latch phase, how long would it take for the output voltage to change to a differential voltage of 2 V?

10.14 What is the minimum power-supply voltage that can be used for the comparator in Fig. 10.28 if the current source and the resistive loads both require 0.4 V across them? You may assume that for the bipolar transistors $V_{CE\text{-}sat} = 0.3$ V.

CHAPTER

11 Sample-and-Hold and Translinear Circuits

This chapter describes some common analog building blocks other than those already presented. Specifically, circuits to realize sample-and-holds, and translinear gain cells and multipliers are discussed. A *sample-and-hold* (S/H) is used to sample an analog signal and to store its value for some length of time. A *translinear gain cell* is commonly used to create an amplifier whose gain can be adjusted through the use of a controlling input current. Finally, a *translinear multiplier* is realized by interconnecting two gain cells and finds a variety of uses in applications such as a modulator (in a communication system) or a phase detector (in a phase-locked loop).

11.1 PERFORMANCE OF SAMPLE-AND-HOLD CIRCUITS

An important analog building block, especially in data-converter systems, is the sample-and-hold circuit. Before proceeding, it is worthwhile to mention that sample-and-hold circuits are also often referred to as "track-and-hold" circuits. Normally, these two terms are synonymous except for a few particular switched-capacitor sample-and-hold circuits that do not have a phase where the output signal is tracking the input signal. Sample-and-hold circuits are necessary components in many data-acquisition systems such as A/D converters. In many cases, the use of a sample-and-hold (at the front of the data converter) can greatly minimize errors due to slightly different delay times in the internal operation of the converter.

Before discussing the basic principles of sample-and-hold circuits, it is necessary to mention some performance parameters used in characterization.

1. The first of these parameters is a *sampling pedestal* or a *hold step*. This is an error that occurs each time a sample-and-hold goes from sample mode to hold mode. During this change in operation, there is always a small error in the voltage being held that makes it different from the input voltage at the time of sampling. Obviously, this error should be as small as possible. *Perhaps more importantly, this error should be signal independent; otherwise it can introduce nonlinear distortion*.
2. Another parameter is a measure of how isolated the sampled signal is from the input signal during hold mode. Ideally, the output voltage will no longer be affected by changes in the input voltage. In reality, there is always some signal feedthrough, usually through parasitic capacitive coupling from the input to the output. In well-designed sample-and-holds, this signal feedthrough can be greatly minimized.
3. A third important parameter is the speed at which a sample-and-hold can track an input signal, when in sample mode. In this mode, a sample-and-hold will have both small-signal and large-signal limitations due to its –3-dB bandwidth and finite slew rate, respectively. Both the –3-dB bandwidth and slew rate should be maximized for high-speed operation.
4. Yet another limitation (somewhat less important in high-speed designs) is the *droop rate* in hold mode. This error is a slow change in output voltage, when in hold mode, caused by effects such as leakage

currents due to the finite base currents of bipolar transistors and reverse-biased junctions. In most CMOS designs, this droop rate is so small it can often be ignored.

5. A fifth limitation is *aperture jitter* or *aperture uncertainty.* This error is the result of the effective sampling time changing from one sampling instance to the next and becomes more pronounced for high-speed signals. Specifically, when high-speed signals are being sampled, the input signal changes rapidly, resulting in small amounts of aperture uncertainty causing the held voltage to be significantly different from the ideal held voltage.

Other performance parameters are also important when realizing sample-and-holds. These include parameters such as dynamic range, linearity, gain, and offset error. Some of the mechanisms whereby these errors arise in particular sample-and-holds are described in the next section using specific S/H examples. In addition, a number of design principles whereby these errors can be minimized are explained.

Key Point: *In addition to usual analog performance metrics such as dynamic range, linearity, etc., sample-and-hold circuits exhibit other nonidealities typically characterized in terms of their sampling pedestal, input isolation, tracking bandwidth, slew rate, droop rate, and aperture uncertainty.*

11.1.1 Testing Sample-and-Holds

Key Point: *The beat test allows a sample-and-hold (or A/D converter) to be characterized at its maximum clock frequency while monitoring only a relatively low-frequency output signal.*

Before describing various S/H architectures, it is appropriate to first describe a popular method for testing S/Hs, referred to as the *beat test*. This test consists of clocking the S/H at its maximum clock frequency and applying a sinusoidal input signal that has a frequency slightly different than the clock frequency.[1] The output of this system is demodulated to a low frequency equal to the difference between the frequency of the clock signal and that of the input signal. This low-frequency signal is then characterized using a spectrum analyzer or by digitizing it using a high-accuracy A/D converter clocked at the difference frequency and then analyzed using a computer. The test setup for a beat test is shown in Fig. 11.1. Example waveforms for the input signal, the sampling signal, and the output signal of the sample-and-hold are shown in Fig. 11.2. Note that this test allows a S/H to be characterized while operating at its maximum clock frequency, yet only a relatively low-frequency output signal must be monitored.

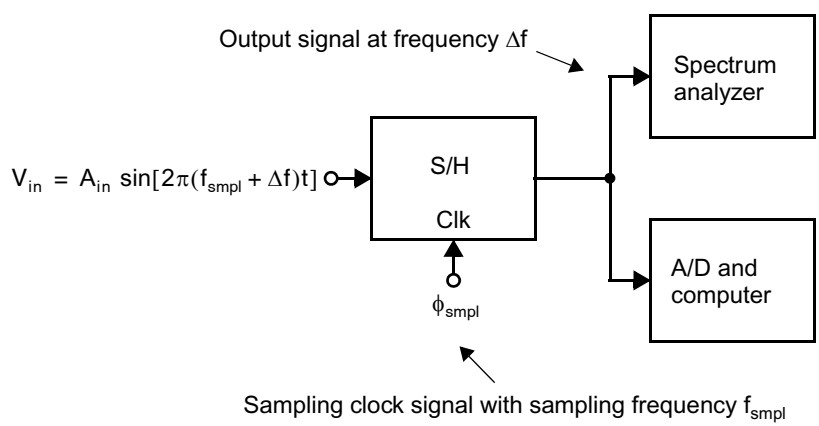


Fig. 11.1 The test setup for characterizing a sample-and-hold using a beat test.

1. If the circuit being tested is, in fact, a track and hold, then the output of the track and hold would be resampled by a second track and hold clocked on the opposite phase as the original.

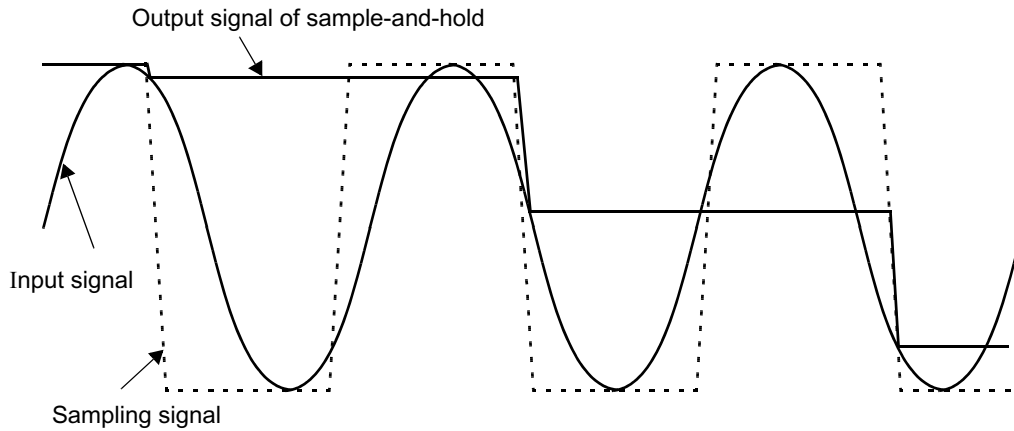


Fig. 11.2 Example waveforms for the test setup of Fig. 11.1.

If a computer is used for analysis, the signal would normally be curve-fitted to a sinusoidal wave at the beat (i.e., difference) frequency. This ideal closest-fit sinusoid would then be subtracted from the measured signal, and the error signal would be analyzed for RMS content and spectral components using a fast-Fourier transform.

It might be mentioned that a very similar test setup can be used for analog-to-digital converters, in which case the output of the A/D converter would be digitally loaded into a computer for analysis.

11.2 MOS SAMPLE-AND-HOLD BASICS

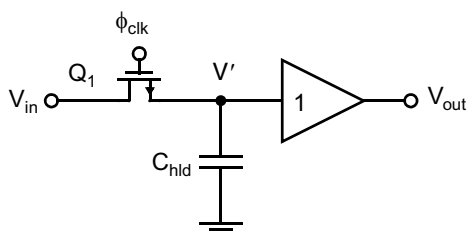


Fig. 11.3 An open-loop track and ho ld realized using MOS technology.

Perhaps the simplest sample-and-hold that can be realized using a CMOS technology is shown in Fig. 11.3. When ϕ_{clk} is high, V' follows V_{in}. When ϕ_{clk} goes low, V' will ideally stay constant from then on, having a value equal to V_{in} at the instance ϕ_{clk} went low. Unfortunately, V' will have a negative going hold step at this time caused by the channel charge of Q_1. When Q_1 turns off, its channel charge must flow out from under its gate into its junctions. Since this charge is negative, it will cause the junction voltages to have negative glitches. If one assumes the source impedance at the node V_{in} is very low, then the glitch at this node will be small and have a very short duration. However, the negative charge that goes to the node with C_{hld} connected to it (i.e., node V') will cause a negative voltage change that is long lasting (until the next time Q_1 turns on again). If clock ϕ_{clk} turns off fast, then the channel charge, Q_{CH}, will flow equally into both junctions [Shieh, 1987] since the channel becomes pinched off at both ends (i.e., next to the junctions) while the charge is flowing out to the two junctions. The charge flowing to the junction labelled V' is therefore given by

$$\Delta Q_{C_{hld}} = \frac{Q_{CH}}{2} = \frac{C_{ox}WLV_{eff\text{-}1}}{2} \tag{11.1}$$

where $V_{eff\text{-}1}$ is given by

$$V_{eff\text{-}1} = V_{GS1} - V_{tn} = V_{DD} - V_{tn} - V_{in} \tag{11.2}$$

Here, V_{in} is the input voltage at the instance Q_1 turns off. It should be noted here that this result assumes that the clock signal, ϕ_{clk}, goes between V_{DD} and the most negative voltage in the circuit.

The change in the voltage V' is found by using the relationship $Q = CV$, resulting in

$$\Delta V' = \frac{\Delta Q_{C\text{-}hld}}{C_{hld}} = -\frac{C_{ox}WLV_{eff\text{-}1}}{2C_{hld}} = -\frac{C_{ox}WL(V_{DD} - V_{tn} - V_{in})}{2C_{hld}} \tag{11.3}$$

Notice that $\Delta V'$ is linearly related to V_{in}, which results in a gain error for the overall sample-and-hold circuit. However, more importantly, $\Delta V'$ is also linearly related to V_{tn}, which is nonlinearly related to the input signal, V_{in}, due to variations in the source-substrate voltage triggering a body effect (assuming the substrate is tied to one of the voltage rails). This nonlinear relationship with V_{in} results in distortion for the overall sample-and-hold circuit.

Key Point: *Charge injection causes a hold step that is nonlinearly related to the input signal, resulting in signal distortion.*

There is also an additional change in V' due to the gate overlap capacitance. Using a derivation similar to that used to find (10.8), we have

$$\Delta V' \cong -\frac{C_{ox}WL_{ov}(V_{DD} - V_{SS})}{C_{hld}} \tag{11.4}$$

where V_{SS} is the most negative voltage in the circuit. This component is usually smaller than that due to the channel charge, and appears simply as an offset since it is signal independent. Therefore, this error component is not typically important since signal-independent offsets can often be removed in most systems. However, it may cause noise problems if care is not taken to ensure that the clock signal, ϕ_{clk}, is relatively noise free. For example, if the clock signal has power-supply noise on it due to being realized with a simple inverter tied to digital power supplies, the power-supply rejection ratio of this circuit might be poor.

EXAMPLE 11.1

Consider the sample-and-hold of Fig. 11.3 with $C_{hld} = 1$ pF, $C_{ox} = 8.5$ fF/(μm)^2, $V_{tn} = 0.45$ V, and $(W/L)_1 = (5\ \mu m/0.4\ \mu m)$. Assume the power supply voltage is 2 V. Find the hold step for V_{in} equal to 1 V, and then repeat for V_{in} equal to 0.5 V. Use these estimates for errors to estimate the dc offset.

Solution

At $V_{in} = 1$ V, using (11.3), we find $\Delta V'(1\text{ V}) = -4.7$ mV. At $V_{in} = 0.5$ V, again using (11.3), we find $\Delta V'(0.5\text{ V}) = -8.9$ mV. The average offset is given by

$$V_{offset\text{-}avg} = \frac{\Delta V'(1\text{ V}) + \Delta V'(0.5\text{ V})}{2} = -6.8\text{ mV} \tag{11.5}$$

which is a reasonable estimate of the amount of dc offset injected in this example.

There have been a number of changes proposed to minimize the signal-dependent hold step. One approach is to replace the n-channel switch by a CMOS transmission gate, as shown in Fig. 11.4. The idea behind this approach is that if the size of the p-channel transistor is taken the same as the n-channel transistor, then the charge injection due to each transistor will cancel when the transmission gate turns off. This result is somewhat true when V_{in} is in the middle region between the power supplies, assuming the clock waveforms are fast and

exactly complementary. Unfortunately, these conditions are seldom possible to achieve in practice. When the finite slopes of the clock waveforms are taken into account, it is seen (as is explained shortly) that the turn-off times of the transistors are signal dependent and this signal dependence causes the n-channel transistor to turn off at different times than the p-channel transistor. Even if we ignore the errors caused by imperfect clock waveforms, when V_{in} is closer to V_{DD}, the charge from the p-channel transistor is greater than that from the n-channel transistor because of its larger effective gate-source voltage resulting in a positive hold step. The opposite happens when the input signal is closer to the negative power supply. Seldom will these two effects have the same magnitude in practice.

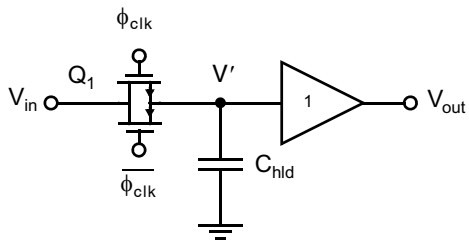


Fig. 11.4 An open-loop track and hold realized using a CMOS transmission gate.

Another modification, often proposed to minimize clock-feedthrough errors, is to add a dummy switch as shown in Fig. 11.5 [McCreary, 1975]. The theory behind this technique is that if the width of Q_2 is taken exactly one-half that of Q_1, and if the clock waveforms are fast, then the charges will cancel. In practice, it is seldom possible to have the clock waveforms change fast enough so that the ideal ratio of widths is exactly one-half. When the ideal ratio is not one-half, it is difficult to make the ratio equal to the optimum required for perfect cancellation. However, when the clock waveforms are fast, this technique usually can minimize the hold pedestal to less than about one-fifth the value it would have without it. For this to be the case, however, it is necessary that the clock of Q_2 changes slightly after that of Q_1. This clock arrangement guarantees that the cancelling charge of Q_2 cannot escape through Q_1 while it is still on.

Key Point: *The effective sampling time is related to the input signal due to the finite rise/fall time of the clock, another nonideality.*

Another source of error for the simple sample-and-hold of Fig. 11.3 is caused by clock waveforms having finite slopes. To understand this error source consider the waveforms for V_{in} and ϕ_{clk} shown in Fig. 11.6. Assume the ideal sampling time is defined to be the negative-going zero-crossing of ϕ_{clk}. The true sampling time is that when the sampling-clock voltage passes through the value when it is one transistor threshold voltage drop above the input voltage. In other words, transistor Q_1 of Fig. 11.3 turns off when ϕ_{clk} is V_{tn} above V_{in}, as shown in Fig. 11.6. Thus, when V_{in} is high, the true sampling time is earlier than the ideal sampling time, whereas when V_{in} is low 0 V, the true sampling time is late.

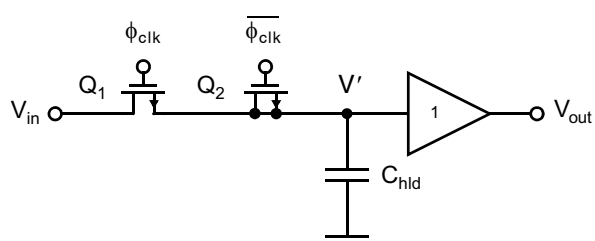


Fig. 11.5 An open-loop track and hold realized using an n-channel switch along with a dummy switch for clock-feedthrough cancellation. The size of Q_2 is one-half that of Q_1.

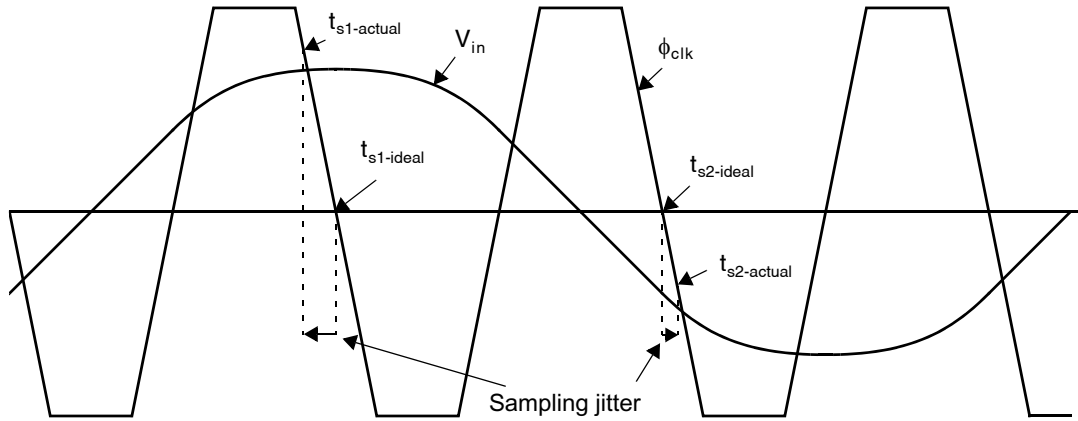


Fig. 11.6 The clock waveforms for V_{in} and ϕ_{clk} used to illustrate how a finite slope for the sampling clock introduces sampling-time jitter.

EXAMPLE 11.2

Consider the S/H circuit of Fig. 11.3, where V_{in} is a 20-MHz band-limited signal with a 2-V_{pp} amplitude. Assume that ϕ_{clk} is a 100-MHz square wave having a peak amplitude of ±2.5 V with linear rise and fall times of 1.5 ns. What is the maximum uncertainty of the sampling time? Assume V_{tn} is 0.8 V.

Solution

First, we note that the slope of ϕ_{clk} is (5 V)/(1.5 ns) = 3.33 V/ns, and the true sampling time is when the clock waveform is 0.8 V greater than the input signal.

For V_{in} equal to 1 V, the sampling transistor will turn off when ϕ_{clk} is 1.8 V, which is (0.7 V)/(3.33 V/ns) = 0.21 ns after the clock first starts to move down.

When V_{in} is –1 V, the sampling time occurs when ϕ_{clk} is –0.2 V, which is 0.81 ns after the clock first starts to move upward.

Therefore, the sampling-time uncertainty is 0.81 – 0.21 = 0.6 ns. Also, assuming the ideal sampling time is 0.75 ns after the clock first starts to move, the sampling jitter is from –0.54 ns to +0.06 ns from the ideal sampling time.

Key Point: *The use of opamps and feedback can provide several performance advantages in S/H circuits, although stability requirements generally limit their speed to less than that of open-loop S/H circuits.*

A more elaborate sample-and-hold circuit is to include an opamp in a feedback loop, as shown in Fig. 11.7. When the clock, ϕ_{clk}, is high, the complete circuit responds similarly to an opamp in a unity-gain feedback configuration. When ϕ_{clk} goes low, the input voltage at that time is stored on C_{hld}, similarly to a simple sample-and-hold. By including an opamp in the feedback loop, the input impedance of the sample-and-hold is greatly increased. Another advantage of this configuration is that even if the unity-gain buffer at the output has an offset voltage, the dc error due to this buffer will be divided by the gain of the input opamp (although the input-offset of the input opamp will remain). Thus, very simple source followers can be used for the output buffer.

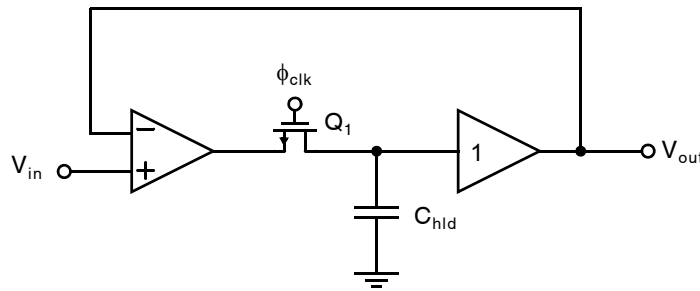


Fig. 11.7 Including an opamp in a feedback loop of a sample-and-hold to increase the input impedance.

A disadvantage of the configuration shown is that the speed of operation can be seriously degraded due to the necessity of guaranteeing that the loop is stable when it is closed. Another source of speed degradation is that when in hold mode, the opamp is open loop, resulting in its output almost certainly saturating at one of the power-supply voltages. When the sample-and-hold next goes back into track mode, it will take some time for the opamp output voltage to slew back to its correct closed-loop value. This slewing time can be greatly minimized by adding two additional transistors as shown in Fig. 11.8. During hold mode, switch Q_2 keeps the output of the first opamp close to the voltage it will need to be at when the T/H goes into track mode. It should also be noted that this track-and-hold configuration still has errors due to charge injection from the switch, Q_1, similar to the simple S/H of Fig. 11.3. The errors due to finite clock rise and fall times are also similar.

A similar, but improved, configuration is shown in Fig. 11.9 [Stafford, 1974; Lim, 1991]. In this configuration, the holding capacitor is not going to ground, but rather is placed in the feedback path of a second opamp. This configuration has a number of desirable features. Perhaps the most important of these is due to the fact that, assuming the second opamp has a large gain, the voltages on both sides of switch Q_1 are very nearly signal independent. Thus, when Q_1 turns off, there will still be charge injection to the left side of C_{hld}, which will cause the output voltage of opamp 2 to have a positive hold step, *but this hold step will be just a dc offset and will be signal independent*. In other words, the charge injection due to Q_1 will cause some dc offset but no distortion. In addition, the sampling time will not change because of the finite slopes of the sampling-clock waveform.

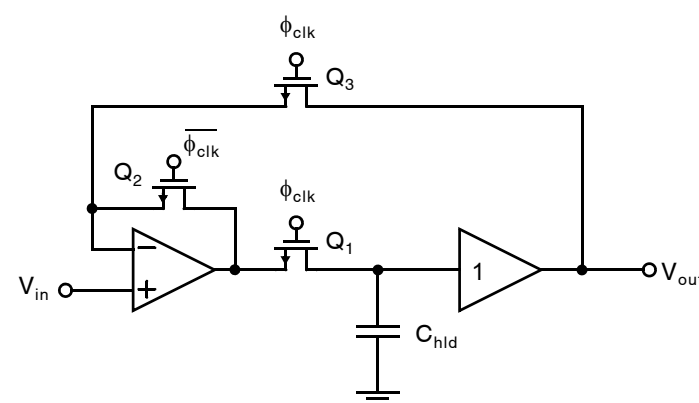


Fig. 11.8 Adding an additional switch to the S/H of Fig. 11.7 to minimize slewing time.

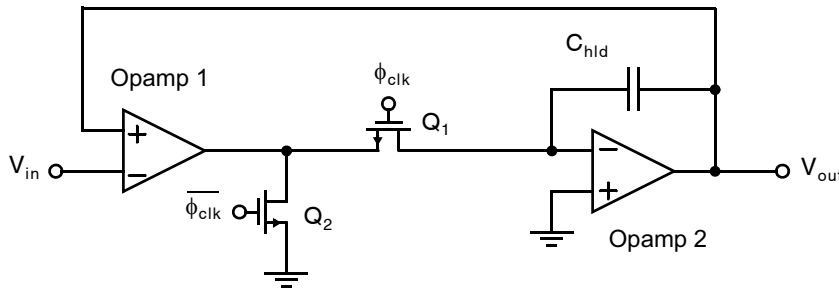


Fig. 11.9 An improved configuration for an S/H as compared to that of Fig. 11.8.

Another advantage is due to the inclusion of Q_2. This switch grounds the output of the first opamp during hold mode. (If the supply voltage is unipolar, the source of Q_2 should instead be connected to a mid-rail dc voltage, as should the positive terminal of Opamp 2.) This grounding keeps the output of the first opamp close to the voltage it must change to when the S/H goes back into track mode. This approach greatly speeds up the time it takes the S/H to return to track mode. The switch also greatly minimizes signal feedthrough when the S/H is in hold mode by grounding the signal path.

A major limitation of this configuration is that the speed will be degraded because of the necessity to guarantee stability in the track mode. This limitation is worsened since there are now two opamps in the loop along with the on resistance of the sampling switch, Q_1, during the closed-loop phase.

Figure 11.10 shows an interesting modification of Fig. 11.9 by including some additional circuitry intended to minimize the hold pedestal and thereby minimize the dc offset [Martin, 1987; Nayebi, 1989]. The basic idea is to match the charge injection into C_{hld} with a similar charge injection into C'_{hld}. Since these capacitors are chosen to have the same size, their voltage changes will match, and the common-mode rejection of the opamp will eliminate the effects of these voltage changes on the output voltage of the S/H. The major limitation of this approach is a second-order effect caused by a mismatch in impedance levels at the left of Q_1 and the bottom of Q_2. Errors due to this mismatch can be minimized by including small capacitors (0.5 pF to 1 pF) between each node and ground. These extra capacitors help keep these nodes at constant voltages while the clocks are turning off, assuming the clock waveforms are fast. It should be mentioned that the configuration of [Nayebi, 1989] was a fully differential design and is a very reasonable choice for many S/H applications.

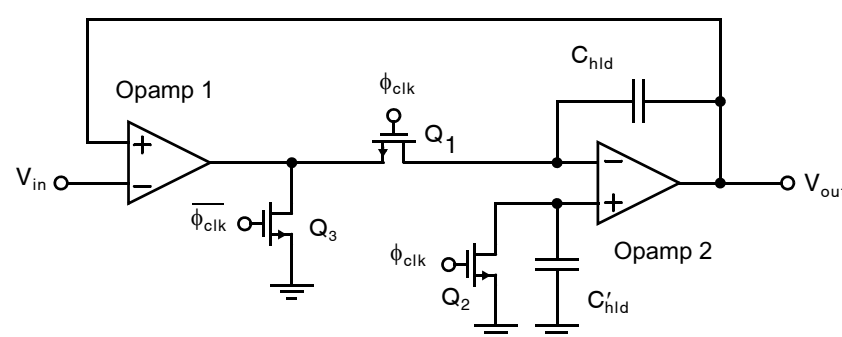


Fig. 11.10 An S/H similar to that of Fig. 11.9, but with clock-feedthrough cancellation circuitry added.

11.3 EXAMPLES OF CMOS S/H CIRCUITS

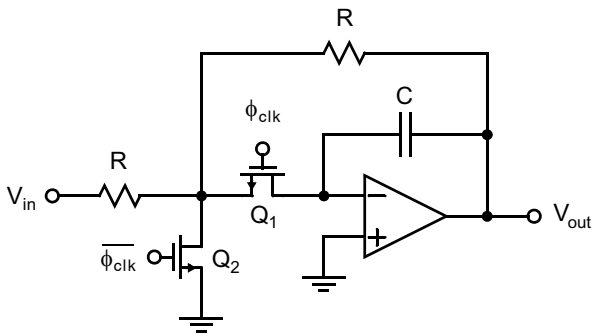


Fig. 11.11 An inverting track and hold.

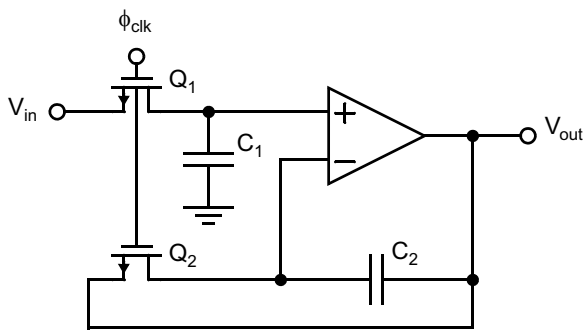


Fig. 11.12 A simple noninverting sample-and-hold with clock-feedthrough cancellation.

In this section, a number of CMOS S/H circuits are described. The optimum choice for a sample-and-hold circuit is highly dependent on the application's requirements and the technology available. Therefore, choosing the best circuit is nontrivial and a matter of ongoing research. For these reasons, it is difficult to definitively state which configurations are superior or inferior.

Before proceeding, it should be noted that many of these circuits could also be realized in a BiCMOS technology, where a speed advantage would be obtained by increasing the unity-gain frequency of any opamps used. For even higher speeds, a BiCMOS circuit can make use of a bipolar diode bridge, as described in Section 11.4.

The S/H circuit shown in Fig. 11.11 [Ishikawa, 1988] is intended to operate at high speeds, assuming a high-speed opamp capable of driving resistive loads is available. Unfortunately, in CMOS technologies, such opamps are difficult to obtain. Furthermore, the required output buffer would not only limit speed, but could also limit the signal swing. In bipolar and BiCMOS technologies, this configuration could be a very viable alternative. When in track mode, Q_1 is on and Q_2 is off, resulting in the sample-and-hold acting as an inverting low-pass circuit having a –3-dB frequency given by $\omega_{-3\,dB} = 1/(RC)$. When Q_1 turns off, the output voltage will remain constant from then on. Since the junctions of Q_1 are always at voltages very close to ground, the clock feedthrough of this sample-and-hold is signal independent. Also, finite clock rise and fall times do not make the sampling time a function of the signal, either. The function of Q_2 is to minimize signal feedthrough when in hold mode and to keep the common node of the resistive network close to the voltage required when the S/H goes back into track mode. In the implementation described in [Ishikawa, 1988], an additional small bypass capacitor was included in parallel to the input resistor. It was reported, and supported by analysis, that this capacitor improved the speed of the S/H.

Another alternative is shown in Fig. 11.12 [Sone, 1993]. This track-and-hold places the opamp into a unity-gain follower during track mode. Also, the positive input terminal of the opamp is connected to V_{in} at this time. When the T/H goes into hold mode, the input signal is stored across C_1, since Q_1 is turned off. At the same time, transistor Q_2 is also turned off. Ideally, the charge feedthrough from Q_1, although signal dependent, will be matched by the charge feedthrough from Q_2, since both are at the same voltage and have the same clock waveforms. For this to be true, the opamp must have a low-impedance output.[2] Furthermore, the signal dependence of the clock signals can be minimized by having clock signals that change above and below the input signal by fixed

2. This can be partially realized using a high-output impedance OTA by having a load capacitor connected to ground, but without having a resistor included in series with the load capacitor (as would be required if lead compensation was desired).

amounts. As will be seen in the next section, this is not too difficult to realize using diode clamps in a bipolar or BiCMOS technology. In a CMOS technology, it is more difficult. Regardless, the topology of Fig. 11.12 is attractive for high-speed CMOS applications that are not too demanding.

Yet another S/H architecture is shown in Fig. 11.13 [Lim, 1991]. Shown in Fig. 11.14(*a*) is the circuit configuration when the S/H is in sample mode. During this mode, the opamp is being reset, and both capacitors are connected between the input voltage and the virtual input of the opamp. Thus the total capacitance that needs to be charged or discharged during this mode is $C_1 + C_2$. During hold mode, the configuration is as shown in Fig. 11.14(*b*). During this mode, the effective hold capacitor is the Miller capacitor given by

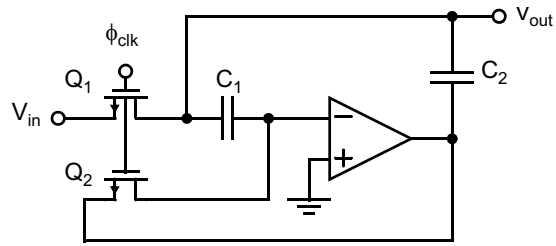


Fig. 11.13 An open-loop architecture but with a Miller holding capacitor.

$$C_{hld-eff} = (1 + A)\left(\frac{C_1 C_2}{C_1 + C_2}\right) \qquad (11.6)$$

which is typically much larger than the capacitance that needs to be charged during the sampling mode. This allows smaller capacitances and therefore smaller switches to be used than would otherwise be the case. Since the voltage changes at the output of the amplifier are very small, it is easy to design the amplifier for very high speed. The sampling switch, Q_1, does inject signal-dependent charge as in other open-loop architectures, but again this is minimized by the large size of the Miller hold capacitance, especially if Q_2 is turned off slightly before Q_1.

At lower frequencies, there are a number of sample-and-hold architectures based on switched-capacitor technology that are often used. These S/Hs tend to be quite accurate but not necessarily fast. A simple example is shown in Fig. 11.15. Note how similar this S/H is to the S/H of Fig. 11.13. During ϕ_1, capacitor C_H is connected between the input signal source and the inverting input of the opamp. At the same time, the inverting input and the output of the opamp are connected together. This causes the voltages at both of these nodes to be equal to the input-offset voltage of the opamp, V_{off}. Thus, C_H is charged up to $V_{in} - V_{off}$. Next, during ϕ_2, the opamp is taken out of reset mode, and the capacitor C_H is connected between the opamp output and its inverting input. This will cause the output voltage to be equal to V_{in}, irrespective of the input-offset voltage of the opamp. Note that this circuit is an example of a sample-and-hold that is not thought of as a track-and-hold circuit. Specifically, during the sample phase, the opamp output voltage will be very close to 0 V (or whatever reference potential is connected to the positive opamp terminal in Fig. 11.15) rather than tracking the input voltage. This places a requirement on the opamp to have a very large slew rate, or the speed is seriously degraded. Also, obviously, the opamp output is invalid during ϕ_1, the sample phase.

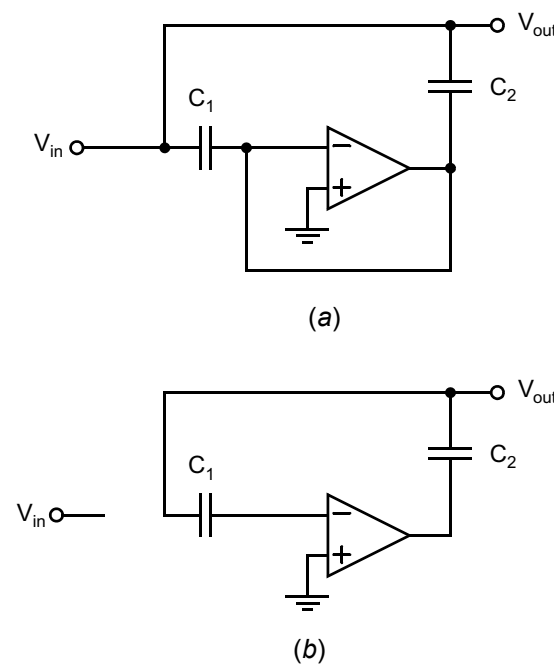


Fig. 11.14 The circuit configurations of the S/H of Fig. 11.13 during (*a*) sample mode and (*b*) hold mode.

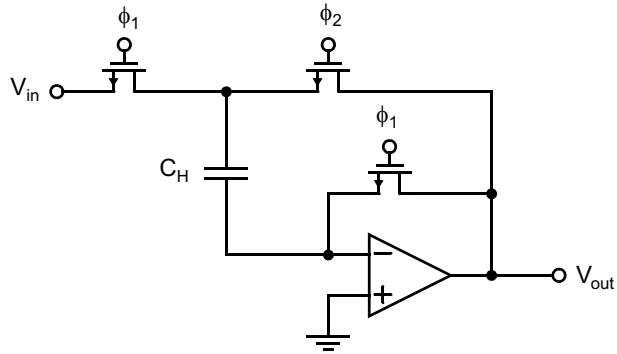


Fig. 11.15 A simple switched-capacitor S/H.

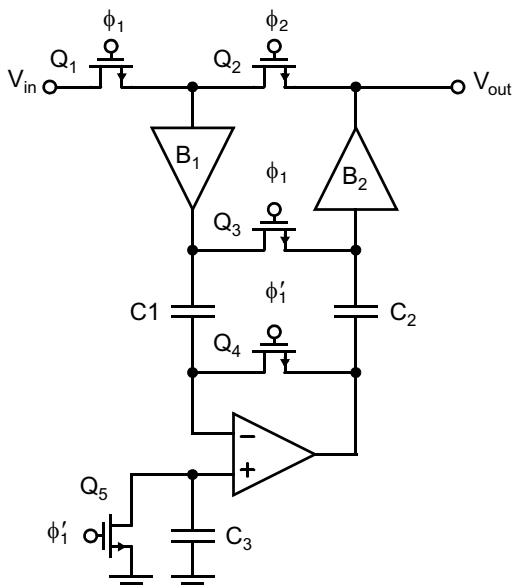


Fig. 11.16 A recycling S/H.

An interesting variation somewhat similar to the S/Hs of Figs. 11.13 and 11.15 adds two unity-gain buffers and two more switches to help improve accuracy. The S/H is shown in Fig. 11.16 [Real, 1991]. During sampling, all switches are on except Q_2. This connects C_1 and C_2 together and charges them to the input voltage. During this time, the input impedance is quite large due to the inclusion of buffer B_1. At the same time, the opamp is being reset and the positive terminal of the opamp is charged to 0 V. Next, Q_4 and Q_5 are turned off, and shortly thereafter Q_1 and Q_3 are turned off. Finally, Q_2 is turned on. This configures the circuit as shown in Fig. 11.17. Note the similarity between this configuration and that of Fig. 11.14(*b*). The unity-gain feedback holds the output during this phase, and also the output impedance of the S/H is very low. Because Q_4 and Q_5 turn off slightly before Q_1 and Q_3, their clock feedthrough is not only signal independent but largely cancels due to the common-mode rejection of the input stage of the amplifier [Martin, 1987]. The charge injection of Q_1 and Q_2 does not affect the output voltage. The charge injection of Q_3 is signal dependent and does affect the output voltage, but its effect is minimized by the loop gain similar to the S/H of Fig. 11.13.

An interesting variation on the S/Hs of Figs. 11.10 and 11.15 is the S/H shown in Fig. 11.18 from [Gatti, 1992]. This T/H has an output that is always valid and incorporates a number of features. During sample mode, the input-offset voltage of the first opamp is stored across C_{OF}. At the same time, C_S is sampling the input voltage, and the second opamp is holding the previously-sampled input voltage. During the next phase, the input-offset voltage of the first opamp is eliminated, and C_S is connected to the output while the feedback loop is enabled. This arrangement causes the output voltage to be equal to the just-sampled input voltage. The clock feedthrough of switch S_2 is cancelled by the clock feedthrough of switch S_3. The clock feedthrough of switch S_5 is cancelled by the clock feedthrough of switch S_6 and the added buffer network. The other switches ideally do not cause appreciable errors due to clock feedthrough. The details concerning the operation of this T/H and the required clocking waveforms are left for the interested reader to consult [Gatti, 1992].

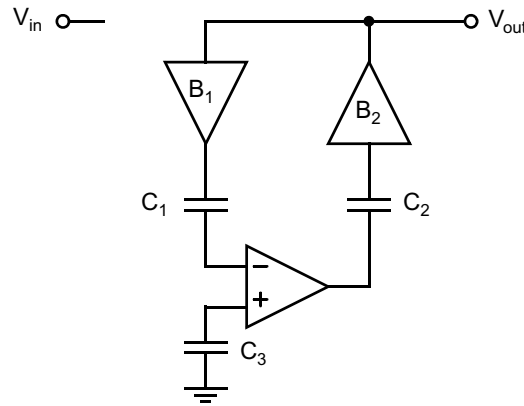


Fig. 11.17 The S/H of Fig. 11.16 when in hold mode.

Yet another example of a switched-capacitor S/H is shown in Fig. 11.19 [Sooch, 1991]. This circuit combines a sample-and-hold with a low-pass filter and was used to convert a signal from the sampled-data domain to the continuous-time domain at the output of a high-quality audio digital-to-analog converter based on oversampling techniques. The low-pass filtering was used to help convert from a single-bit high-frequency signal to a decimated lower-frequency multi-bit signal. During ϕ_1, C_1 is connected between the input voltage and ground. Next, during ϕ_2, it is connected between the output and the inverting input of the opamp. At this time, its charge is shared with the charge being stored across C_2, which came from previous samples. For low-frequency signals, the output voltage will equal the input voltage in the steady state. Higher frequency changes in the input signal will be low-pass filtered. For C_2 greater than C_1 by at least a few times, it can be shown using switched-capacitor analysis techniques [Martin, 1980] that the –3-dB frequency of the low-pass filter is approximately given by

$$f_{-3\,dB} \cong \frac{1}{2\pi}\frac{C_1}{C_2} f_{clk} \tag{11.7}$$

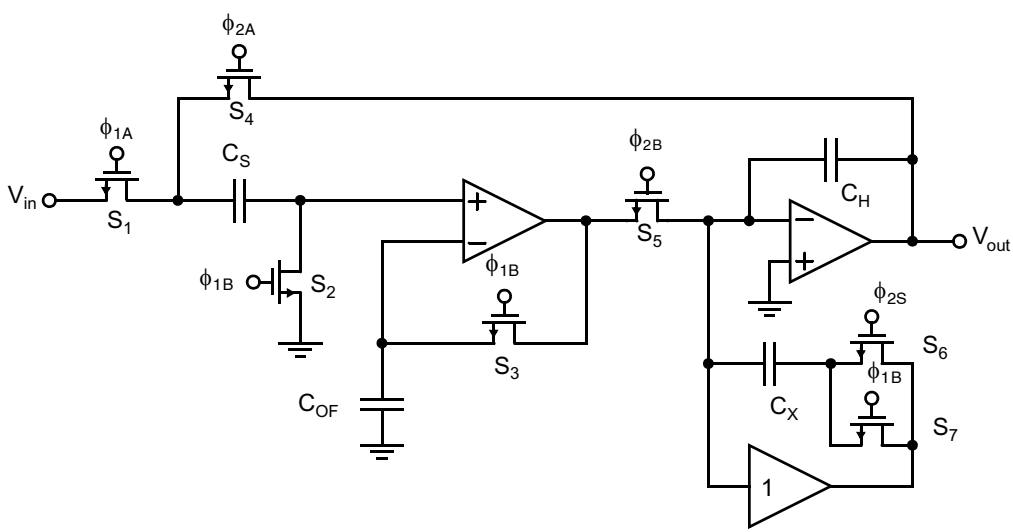


Fig. 11.18 A switched-capacitor S/H.

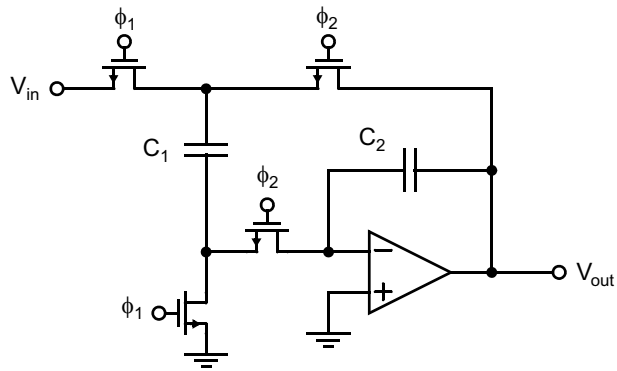


Fig. 11.19 A switched-capacitor sample-and-hold and low-pass filter.

where f_{clk} is the clock frequency. While this example does not eliminate the effects of opamp finite-input voltage offsets, unlike the previous two examples, it does have the advantage that its output is always valid.

The switched-capacitor sample-hold circuits just shown (except for that of Fig. 11.18) do have dc offsets due to clock feedthrough from the switches connected to the inverting inputs of the opamps. However, these offsets are largely signal independent assuming the opamp has a reasonable gain. Also, it is possible to minimize these offsets by including a clock-feedthrough cancellation network connected to the positive input of the opamp, similar to what is described in [Martin, 1987]. Alternatively, fully differential sample-and-holds can be used, which also helps minimize the effects of clock feedthrough.

11.4 BIPOLAR AND BICMOS SAMPLE-AND-HOLDS

In this section, bipolar and BiCMOS sample-and-hold circuits are described. It should be noted that BiCMOS sample-and-hold circuits are often realized using the CMOS techniques previously described. Here, we look at BiCMOS circuits more closely related to their bipolar counterparts.

Most of the early integrated bipolar sample-and-holds were based on using a *diode-bridge* switch, as shown in Fig. 11.20 [Erdi, 1978]. In track mode the bias current, I_B, is turned on and evenly divides between D_1 and D_2, as well as between D_3 and D_4. This current division results in a low-impedance path between the input and the output with a small-signal impedance of r_d, where

$$r_d = \frac{V_T}{I_{D1}} = \frac{V_T}{I_B/2} \tag{11.8}$$

In hold mode, both current sources labelled I_B are turned off, leaving the hold capacitor, C_{hld}, open circuited.

A more practical realization of a diode-bridge track-and-hold is shown in Fig. 11.21 [Matsuzawa, 1990]. During track mode, V_{trk} is at a higher voltage than V_{hld}, and Q_2 will be on and conducting $2I_B$. Half of this current comes from the current source connected to the node V_2. The other half comes from the current source connected to node V_1 after going through the diode bridge. Thus, during track mode, the diode bridge is low impedance, and the node V_3 will track the input signal. The output voltage will also track this node through the unity-gain output buffer. Also, during this phase, the diodes D_5 and D_6 are reverse biased.

When the track-and-hold is placed in hold mode, Q_2 is turned off and Q_1 is turned on. Since Q_2 is turned off, the current from the source connected to the node V_2 now flows through the diode D_6, turning it on. This causes

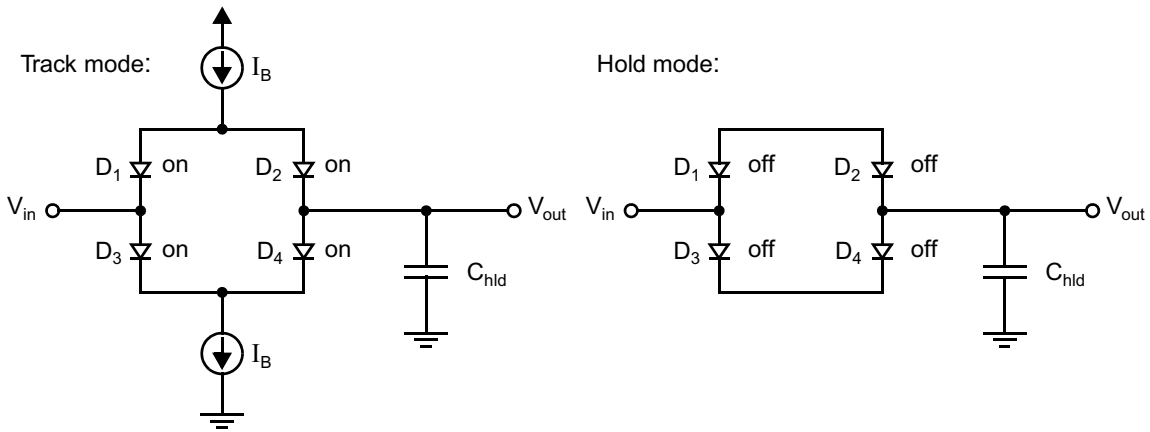


Fig. 11.20 A bipolar track and hold based on using a diode bridge.

V_2 to be one diode drop above V_{out}, which causes diodes D_3 and D_4 to become reverse biased. At the same time, Q_1 will be conducting current $2I_B$. Half of this will come from the current source connected to node V_1. The other half comes from V_{out} through diode D_5, causing it to become forward biased. Therefore, V_1 will be one diode drop below V_{out}, and diodes D_1 and D_2 will be reverse biased. Thus, all the diodes in the diode bridge are reverse biased, and the node V_3 is disconnected from the input signal.

The circuit of Fig. 11.21 has a number of advantages compared to the simple circuit of Fig. 11.20. Perhaps the most important of these is that the clock feedthrough is signal independent. This independence occurs because when the diode bridge turns off, the voltage changes of nodes V_1 and V_2 are from a voltage closely equal to V_{in}, to a voltage

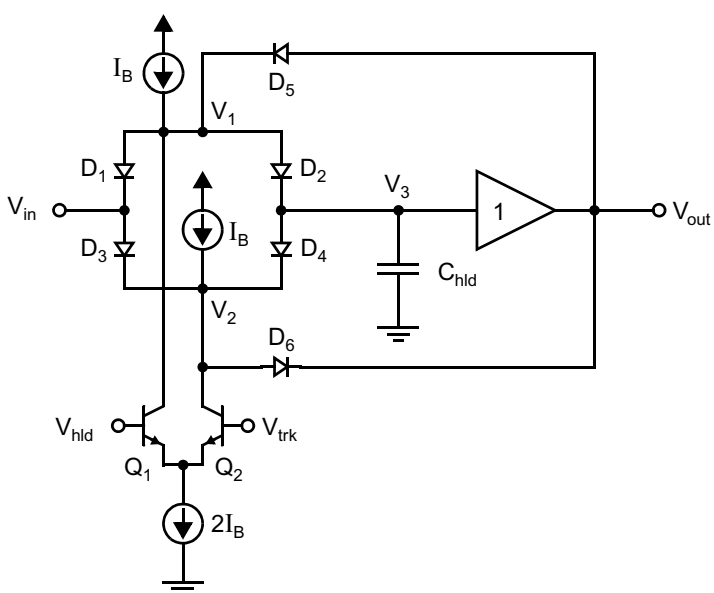


Fig. 11.21 An improved example of a diode-bridge track and hold.

one diode drop below V_{in}, and a voltage one diode drop above V_{in}, respectively, regardless of the value of V_{in}. Thus, the charge from turning the diode bridge off, and changing the reverse-bias voltage of the diodes in the bridge, is signal independent. Furthermore, the charge from D_1 and D_2 will (to a first-order degree) cancel the charge from D_3 and D_4, assuming fast clock waveforms, which also minimizes the effects of charge injection. It should be mentioned that in many implementations, the diodes would all be Schottky diodes, which don't have minority carrier charge storage when they are on, and, thus, turn off faster than regular diodes and have smaller hold steps. A second advantage of Fig. 11.21 is that the sample-time uncertainty caused by the finite slope of the clock waveforms is greatly minimized. Finally, a third advantage of Fig. 11.21 is that when in hold mode, the signal feedthrough is substantially smaller. When the simple circuit of Fig. 11.20 is in hold mode, the diodes are turned off but are floating. The input signal is then capacitively coupled through the junction capacitance of these reverse-biased diodes to the hold capacitor. The magnitude of this signal feedthrough is easily calculated using the capacitor-divider formula. This signal feedthrough is minimized in the improved circuit of Fig. 11.21, because during hold mode, nodes V_1 and V_2 are connected to V_{out} through low-impedance paths consisting of conducting diodes D_5 and D_6, respectively. This effectively shields the node V_3 from the input signal during hold mode.

In the BiCMOS realization presented in [Matsuzawa, 1990], the input transistors to the unity-gain buffer were p-channel MOS transistors that minimized the droop on the hold capacitor during hold-mode, which would normally be caused by finite-based currents of bipolar input transistors.

It should be noted in Fig. 11.21 that the maximum input-signal excursions during hold mode are limited to less than two diode drops above or below the input signal voltage at the time the track and hold went into hold mode. Otherwise, one of the diodes D_1 or D_3 will become forward biased, which will tend to clamp the input signal. For high-speed track and holds, this input amplitude restriction is not usually a serious limitation as peak signals are usually limited to be less than one volt. Furthermore, the input signal range can be increased by replacing each of D_5 and D_6 with two series diodes. This would increase the maximum allowable signal-swing excursion to three diode drops.

Key Point: *S/Hs suffer from voltage droop in hold mode when bipolar transistors draw base currents from the hold capacitors. The droop can be reduced by using MOSFETs instead, or by using a differential signal path so that both sides exhibit similar droop.*

One of the major limitations of bipolar S/Hs is a droop in the voltage being stored across a capacitor during hold mode, caused by the finite base currents of bipolar transistors. One of the major ways of minimizing this droop is to use fully differential approaches where the droop is the same in both paths, and the differential signal is, to a first-order approximation, unaffected. An example of a fully differential diode-bridge-based T/H is given in [Colleran, 1993]. In this example, the current sources were realized using resistors, and additional circuitry was also included to minimize errors introduced by these resistors.

An interesting BiCMOS S/H is shown in Fig. 11.22 [Wakayama, 1992]. This S/H uses bipolar switches that appear similar to the diode bridge switches, but operate differently. Note that the direction of diodes D_2 and D_4 have been reversed compared to the circuit of Fig. 11.20. When V_{trk1} is greater than V_{hld1}, then Q_3 will be conducting $3I_B$. I_B of this will come from M_5 through Q_1, I_B of this will come from M_6, and I_B of this will come from M_7 through D_2 and D_1. Since both Q_1 and D_1 are conducting, V_{out1} will equal V_{in}. Diodes D_3 and D_4 are reverse biased and V_{out2} is isolated from V_{in}. The ratios of the currents in M_2, M_3, and M_4 are accurately determined by the feedback loop including M_1, M_2, M_5, and the unity-gain buffer. When V_{trk1} is less than V_{hld1}, then Q_4 will be conducting $3I_B$. This will cause Q_1, D_1, and D_2 to turn off and Q_2, D_3, and D_4 to turn on. In this phase, V_{out1} will be isolated from V_{in}, and V_{out2} will track V_{in}. Notice that the input is buffered from the holding capacitors by emitter followers (either Q_1 or Q_2) and that the load current on the input source is constant (and equal to the base current of the emitter followers). Also, when one output is in track mode, the other is in hold mode. This arrangement can possibly be used to effectively double a data rate. Finally, as is explained in [Wakayama, 1992], during slewing the current available to charge or discharge the holding capacitors is dynamically increased to $3I_B$ rather than the quiescent I_B.

The BiCMOS S/H shown in Fig. 11.23 is referred to as a switched emitter-follower [Vorenkamp, 1992]. The degenerated differential pair $Q_{1,2}$ serves as an input buffer. When in track mode, the tail current I_2 is steered towards $Q_{4,5}$; this turns on transistors $Q_{6,7}$ which both serve as emitter followers. The hold capacitor C_{hld} is

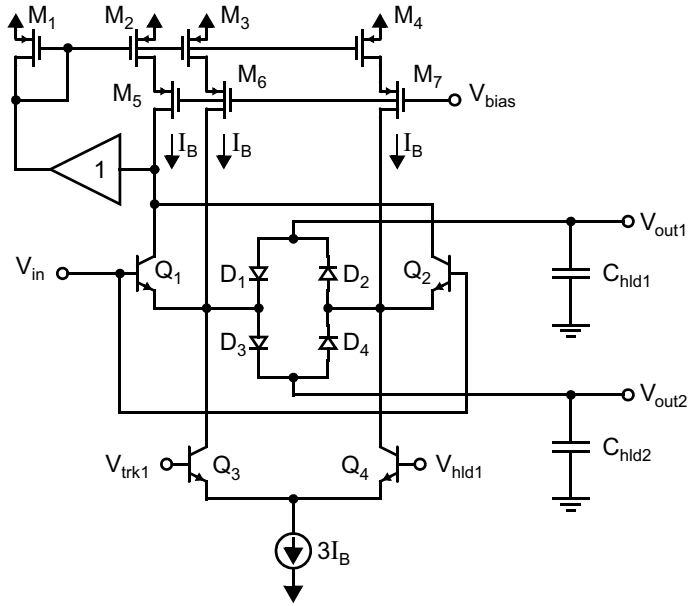


Fig. 11.22 A BiCMOS sample-and-hold.

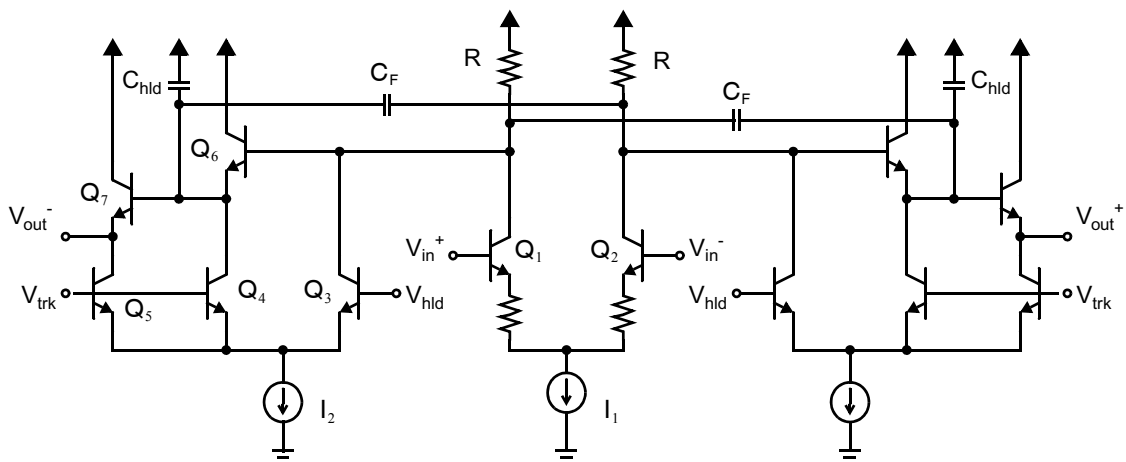


Fig. 11.23 A high-speed bipolar track-and-hold circuit [Vorenkamp, 1992].

therefore driven by a low-impedance buffer. When the circuit switches to hold mode, all of the tail current I_2 is directed away from $Q_{4,5}$, which turn off the emitter followers $Q_{6,7}$ and leave the hold capacitor. Since the current will switch from Q_3 to $Q_{4,5}$ over a very narrow range of differential voltage at $V_{trk} - V_{hld}$, the switching can occur quickly even with a low-swing clock. Also, in hold mode all of I_2 is directed through Q_3 and from there through the resistor R. The resulting additional voltage drop across R reduces the base voltage of Q_6, helping to turn it off quickly. Unfortunately, even when Q_6 is off in hold mode, its finite base-emitter capacitance C_{be6} provides a path for signal feedthrough in hold mode. For this reason, $C_F \approx C_{be6}$ is introduced to provide signal feedthrough of the

same amplitude but opposite polarity, cancelling the feedthrough through C_{be6}. The circuit operates open-loop and can therefore be very fast.

An alternate BiCMOS S/H is described in [Sone, 1993], which makes use of the basic architecture shown in Fig. 11.12, except the switches are realized using bipolar switches. There have been a number of other alternative bipolar and BiCMOS track and holds that have been reported in the literature. Some alternative and interesting BiCMOS examples include [Fernandes, 1989; Robertson, 1990; and Real, 1991]. Some alternative and interesting bipolar examples include [Moraveji, 1991] and [Petschacher, 1990]. The interested reader is referred to the applicable references.

11.5 TRANSLINEAR GAIN CELL

An important analog building block for bipolar circuits is the translinear gain cell[3] [Gilbert, 1968a; Gilbert, 1990]. These gain cells are useful when building circuits requiring varying gain, such as in adjustable continuous-time filters and voltage-controlled amplifiers. They are also useful in realizing four-quadrant analog multipliers, as will be seen in the next section.

A translinear gain cell is shown in Fig. 11.24. It has differential inputs and outputs that are current signals. The first stage simply consists of pn junction loads, which in this case are realized by the base-emitter junctions of Q_1 and Q_2. The second stage is a differential pair. To analyze this cell, we recall the relationship between the base-emitter voltage and collector current,

$$I_C = I_S e^{(v_{be}/V_T)} \quad \text{or} \quad v_{be} = V_T \ln\left(\frac{I_C}{I_S}\right) \tag{11.9}$$

Thus, ignoring the base currents of Q_3 and Q_4, we see that the voltage at the emitter of Q_1 is given by

$$v_{e1} = (-V_T)\ln\left(\frac{I_1 + i_{o1}}{I_{S1}/\alpha}\right) \tag{11.10}$$

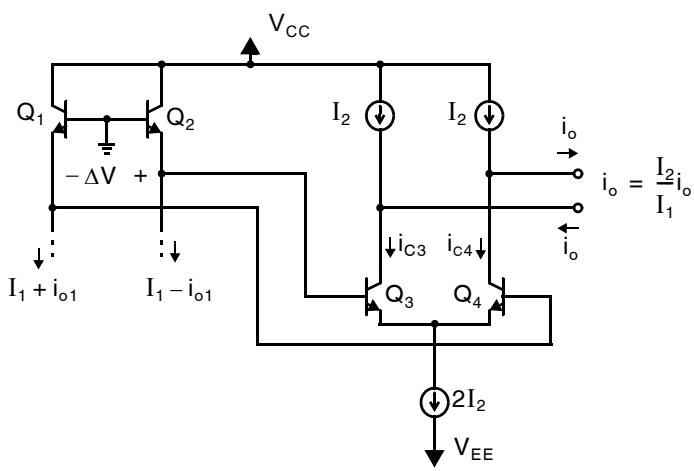


Fig. 11.24 A four-transistor gain cell. The output current, i_o, is a scaled version of i_{o1}.

3. These four transistor gain cells are also commonly called two-quadrant multipliers or Gilbert gain cells.

where $\alpha = \beta/(\beta + 1)$. For Q_2, we have

$$v_{e2} = (-V_T)\ln\left(\frac{I_1 - i_{o1}}{I_{S2}/\alpha}\right) \tag{11.11}$$

Now, assuming Q_1 and Q_2 are matched such that $I_{S1} = I_{S2}$, and defining ΔV to be the difference between v_{e2} and v_{e1}, we have

$$\Delta V \equiv v_{e2} - v_{e1} = V_T \ln\left(\frac{I_1 + i_{o1}}{I_1 - i_{o1}}\right) \tag{11.12}$$

Note that this voltage difference, ΔV, does not depend on the scale current, I_S.

To analyze the current output from the differential pair of Q_3 and Q_4, we recall from Section 3.10 that the large-signal emitter currents through a differential pair are given by

$$i_{E3} = \frac{2I_2}{1 + e^{(-\Delta V)/V_T}} \quad \text{and} \quad i_{E4} = \frac{2I_2}{1 + e^{(\Delta V)/V_T}} \tag{11.13}$$

Therefore, the collector current, i_{C3}, is equal to

$$i_{C3} = \alpha i_{E3} = \frac{\alpha 2I_2}{1 + \left(\dfrac{I_1 - i_{o1}}{I_1 + i_{o1}}\right)} = \alpha I_2 + \alpha\frac{I_2}{I_1} i_{o1} \tag{11.14}$$

and for $\alpha = 1$, we have

$$i_{C3} = I_2 + \frac{I_2}{I_1} i_{o1} \tag{11.15}$$

In a similar manner, i_{C4} is equal to

$$i_{C4} = I_2 - \frac{I_2}{I_1} i_{o1} \tag{11.16}$$

Finally, the output current, i_o, is given by

$$i_o = i_{C3} - I_2 = I_2 - i_{C4} = \frac{I_2}{I_1} i_{o1} \tag{11.17}$$

In other words, the output current is equal to a scaled version of the input current where the scaling factor is determined by the ratio of two dc bias currents, I_2 and I_1.

An important point to note here is that not only is the difference between i_{C3} and i_{C4} linearly related to the input current, *but also each of* i_{C3} *and* i_{C4} *is also linearly related* (except for the dc bias current). This result is important since it implies that even if the difference between these two currents is not taken precisely, it does not affect the linearity of the overall circuit. Furthermore, note that the gain and linearity are not dependent on any transistor parameters such as β or I_S. However, one should be aware that there are mechanisms that will create distortion in a gain cell, such as mismatches between transistors and the occurrence of linear resistances in the pairs Q_1, Q_2 and Q_3, Q_4.

> ***Key Point:*** *Translinear gain cells are important blocks in analog bipolar design, capable of providing very linear variable gains and analog multipliers.*

EXAMPLE 11.3

Consider the gain cell shown in Fig. 11.24 where I_1 is fixed to a bias current of 200 μA, while I_2 varies such that the gain changes from 0.5 to 4. What are the extreme values of I_2 needed? Also, what is the maximum peak current that should be used for i_{o1}, assuming all transistors should remain in the active region with at least 20 μA of emitter current?

Solution

Making use of (11.17), when a gain of 0.5 is needed, I_2 should be one-half of I_1, which is 100 μA. When a gain of 4 is needed, I_2 should be four times that of I_1, which is 800 μA. Therefore, the extreme values of I_2 are 100 μA and 800 μA.

The maximum peak current of i_{o1} is found by noting that the emitter currents of either transistors Q_1 or Q_2 should not go below 20 μA. Since they are biased with 200 μA each, the peak current of i_{o1} should be less than 180 μA. Note that if I_1 were varied while I_2 remained fixed, the maximum peak level of i_{o1} would change as I_1 was varied.

11.6 TRANSLINEAR MULTIPLIER

A translinear multiplier[4] is one approach for realizing a four-quadrant multiplier circuit. Multiplier circuits are useful in a variety of applications, such as frequency modulators in communication systems or phase detectors in phase-locked loops. They rely upon the hyperbolic tangent voltage-current relationship of the bipolar differential pair, repeated here from (8.81)

$$I_{C2} - I_{C1} = \alpha I_{EE} \tanh(V_{id}/2V_T) \tag{11.18}$$

Key Point: *Analog multiplication can be performed by superimposing an input signal on the tail current of a bipolar differential pair. If both of the inputs to be multiplied are differential, two differential pairs are needed.*

Hence, the differential output current is linearly related to the emitter current I_{EE}. The principle of operation is to precede the differential pair input with an inverse hyperbolic tangent nonlinearity (also provided by bipolar transistors), and then to superimpose the second input onto the differential pair's tail current, resulting in a linear multiplication of the inputs. If both of the inputs to be multiplied are differential, two differential pairs are used.

One implementation of a translinear multiplier using two translinear gain cells is the six-transistor circuit in Fig. 11.25 [Gilbert, 1968b]. We can analyze this multiplier by making use of the results for the gain-cell circuit just discussed preceding. Note that the preceding multiplier circuit is effectively two gain-cell circuits with their output signals cross-coupled. Specifically, one gain cell consists of the two differential pairs Q_1, Q_2 and Q_5, Q_6, while the other gain cell consists of the differential pair Q_3, Q_4 and makes use of the same voltage generating pair, Q_5, Q_6. Thus, recognizing the similarity of this circuit with that of Fig. 11.24 and making use of (11.15) and (11.16), we find the following four collector currents:

$$i_{C1} = \frac{(I_2 + i_2)}{2} + \frac{(I_2 + i_2)}{2I_1} i_1 \tag{11.19}$$

4. A translinear multiplier is also commonly called a Gilbert multiplier or an analog multiplier.

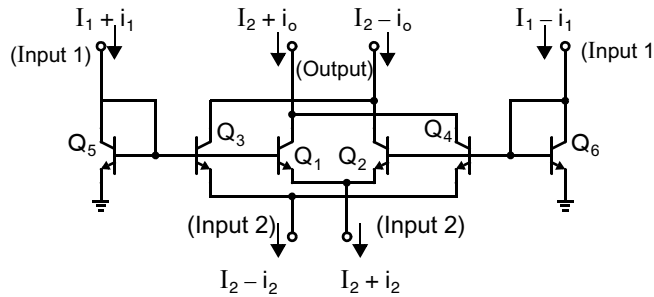


Fig. 11.25 A six-transistor translinear multiplier.

$$i_{C2} = \frac{(I_2 + i_2)}{2} - \frac{(I_2 + i_2)}{2I_1} i_1 \tag{11.20}$$

$$i_{C3} = \frac{(I_2 - i_2)}{2} + \frac{(I_2 - i_2)}{2I_1} i_1 \tag{11.21}$$

$$i_{C4} = \frac{(I_2 - i_2)}{2} - \frac{(I_2 - i_2)}{2I_1} i_1 \tag{11.22}$$

Note here that we have made the substitutions of signal currents, $I_2 \pm i_2$, in what was simply a bias current of $2I_2$ in Fig. 11.24. In addition, once again these four collector currents are linear in i_2 if i_1 remains a constant value and linear in i_1 if i_2 remains constant. To find the output currents, we combine collector currents to obtain

$$I_2 + i_o \equiv i_{C1} + i_{C4} = I_2 + \frac{i_1 i_2}{I_1} \tag{11.23}$$

and

$$I_2 - i_o \equiv i_{C2} + i_{C3} = I_2 - \frac{i_1 i_2}{I_1} \tag{11.24}$$

and taking the difference of these two equations results in

$$i_o = \frac{i_1 i_2}{I_1} \tag{11.25}$$

Thus, we see that the output signals are linear with respect to i_2 and can be scaled by the ratio i_1 / I_1. Also, note that i_1 can be made positive or negative (as it is the amount of the bias current deviation from I_1), and so the output current, i_o, is a scaled positive or negative version of i_2. A similar result occurs with respect to the output signal and i_1. Thus, this multiplier is called a four-quadrant multiplier. Recall that a gain cell can only scale i_1 with respect to the ratio of two positive bias currents and is therefore only a two-quadrant multiplier.

If voltage inputs, rather than current inputs, are desired, then one can precede the multiplier with voltage-to-current converters. The simplest of these are differential pairs having emitter degeneration, as shown in Fig. 11.26. More linear, but more complicated, voltage-to-current converters include local feedback to eliminate distortions due to the modulation of the base-emitter voltages. Examples of these circuits are given in Chapter 12, where continuous-time filters are discussed.

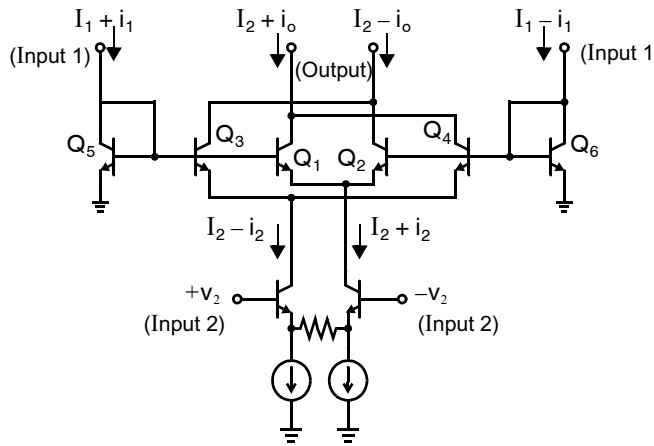


Fig. 11.26 The six-transistor translinear multiplier with a degenerated input differential pair to perform voltage-to-current conversion of input 2.

EXAMPLE 11.4

Consider the translinear multiplier circuit shown in Fig. 11.25, where I_1 and I_2 are fixed bias currents of 100 μA and 200 μA, respectively. If i_1 is the gain-controlling current, what is that maximum gain for i_o/i_2, assuming all transistors should remain in the active region with at least 20 μA of emitter current?

Solution

To maintain 20 μA of current through Q_5 or Q_6, the magnitude of i_1 should be limited to 80 μA. As seen from (11.25), such a limit results in the peak gains being ±0.8. A gain of zero would occur if collector currents of Q_5 and Q_6 were equal.

Note that this multiplier circuit cannot get gain from i_2 to i_o, while gain is possible if the input signal is taken to be i_1.

The translinear multiplier may also be adapted to implementation in CMOS. The effort is challenging because MOSFETs do not provide an exponential voltage-current relationship in active mode, so for high accuracy compensating nonlinearities are introduced. See, for example, [Soo, 1982].

11.7 KEY POINTS

- In addition to usual analog performance metrics such as dynamic range, linearity, etc., sample-and-hold circuits exhibit other nonidealities typically characterized in terms of their sampling pedestal, input isolation, tracking bandwidth, slew rate, droop rate, and aperture uncertainty. [p. 445]
- The beat test allows a sample-and-hold (or A/D converter) to be characterized at its maximum clock frequency while monitoring only a relatively low-frequency output signal. [p. 445]
- Charge injection causes a hold step that is nonlinearly related to the input signal, resulting in signal distortion. [p. 447]

- The effective sampling time is related to the input signal due to the finite rise/fall time of the clock, another non-ideality. [p. 448]
- The use of opamps and feedback can provide several performance advantages in S/H circuits, although stability requirements generally limit their speed to less than that of open-loop S/H circuits. [p. 449]
- S/Hs suffer from voltage droop in hold mode when bipolar transistors draw base currents from the hold capacitors. The droop can be reduced by using MOSFETs instead, or by using a differential signal path so that both sides exhibit similar droop. [p. 458]
- Translinear gain cells are important blocks in analog bipolar design, capable of providing very linear variable gains and analog multipliers. [p. 461]
- Analog multiplication can be performed by superimposing an input signal on the tail current of a bipolar differential pair. If both of the inputs to be multiplied are differential, two differential pairs are needed. [p. 462]

11.8 REFERENCES

W. Colleran and A. Abidi. "A 10-b, 75-MHz Two-Stage Pipelined Bipolar A/D Converter," *IEEE J. of Solid-State Circuits,* Vol. 28, no. 12, pp. 1187–1199, December 1993.

G. Erdi and P. R. Henneuse. "A Precision FET-Less Sample-and-Hold with High Charge-to-Droop Current Ratio," *IEEE J. of Solid-State Circuits,* Vol. 13, no. 6, pp. 864–873, December 1978.

J. Fernandes, S. Miller, M. Mallison, and G. Miller. "A 14-Bit 10-μs Subranging A/D Converter with S/H," *IEEE J. of Solid-State Circuits*, Vol. 24, no. 6, pp. 1485–1491, December 1989.

U. Gatti, F. Maloberti, and G. Palmisano. "An Accurate CMOS Sample-and-Hold Circuit," *IEEE J. of Solid-State Circuits,* Vol. 27, no. 1, pp. 120–122, January 1992.

B. Gilbert. "A New Wide-Band Amplifier Technique," *IEEE J. of Solid-State Circuits,* Vol. 3, pp. 353–365, December 1968a.

B. Gilbert. "A Precise Four-Quadrant Multiplier with Subnanosecond Response," *IEEE J. of Solid-State Circuits,* Vol. 3, pp. 365–373, December 1968b.

B. Gilbert. "Current-Mode Circuits from a Translinear Viewpoint: A Tutorial," *Analog IC Design: The Current-Mode Approach,* ed. C. Toumazou, F. J. Lidgey, and D. G. Haigh, Peter Peregrinus, London, 1990.

M. Ishikawa and T. Tsukahara. "An 8-bit 50-MHz CMOS Subranging A/D Converter with Pipelined Wide-Band S/H," *IEEE J. of Solid-State Circuits*, Vol. 23, no. 6, pp. 1309–1315, December 1988.

P. Lim and B. Wooley. "A High-Speed Sample-and-Hold Technique Using a Miller Hold Capacitance," *IEEE J. of Solid-State Circuits,* Vol. 26, no. 4, pp. 643–651, April 1991.

K. Martin. "Improved Circuits for the Realization of Switched-Capacitor Filters," *IEEE Trans. Circuits and Systems,* Vol. CAS-27, no. 4, pp. 237–244, April 1980.

K. Martin. "New Clock-Feedthrough Cancellation Technique for Analog MOS Switched-Capacitor," *Elctron. Lett.,* no. 18, 1992.

K. Martin, L. Ozcolak, Y. S. Lee, and G. C. Temes. "A Differential Switched-Capacitor Amplifier," *IEEE J. of Solid-State Circuits,* Vol. SC-22, no. 1, pp. 104–106, February 1987.

A. Matsuzawa, et al. "A 10-b 30-MHz Two-Step Parallel BiCMOS ADC with Internal S/H," *IEEE Int. Solid-State Circuits Conf.*, pp. 162–163, February 1990.

J. McCreary and P. R. Gray. "All-MOS Charge Redistribution Analog-to-Digital Conversion Techniques—Part 1," *IEEE J. of Solid-State Circuits*, Vol. SC-10, pp. 371–379, December 1975.

F. Moraveji. "A 14-b, 150-ns Sample-and-Hold Amplifier with Low Hold Step," *IEEE Int. Solid-State Circuits Conf.*, pp. 166–167, February 1991.

M. Nayebi and B. Wooley. "A 10-bit Video BiCMOS Track-and-Hold Amplifier," *IEEE J. of Solid-State Circuits*, Vol. 24, no. 6, pp 1507–1516, December 1989.

R. Petschacher, B. Zojer, B. Astegher, H. Jessner, and A. Lechner. "A 10-b 75MSPS Subranging A/D Converter with Integrated Sample-and-Hold," *IEEE J. of Solid-State Circuits*, Vol. 25, no. 6, pp. 1339–1346, December 1990.

P. Real and D. Mercer. "A 14-b Linear, 250ns Sample-and-Hold Subsystem with Self-Correction," *IEEE Int. Solid-State Circuits Conf.*, pp. 164–165, February 1991.

D. Robertson, P. Real, and C. Mangelsdorf. "A Wideband 10-bit 20Msps Pipelined ADC Using Current-Mode Signals," *IEEE Int. Solid-State Circuits Conf.*, pp. 160–161, February 1990.

J. Shieh, M. Patil, and B. L. Sheu. "Measurement and Analysis of Charge Injection in MOS Analog Switches," *IEEE J. of Solid-State Circuits*, Vol. 22, no. 2, pp. 277–281, April 1987.

K. Sone, Y. Nishida, and N. Nakadai. "A 10-b 100-Msample/s Pipelined Subranging BiCMOS ADC," *IEEE J. of Solid-State Circuits,* Vol. 28, no. 12, pp. 1180–1186, December 1993.

D. C. Soo and R. G. Meyer. "A four-quadrant NMOS analog multiplier," Solid-State Circuits, *IEEE J. of Solid-State Circuits,* Vol.17, pp. 1174–1178, Dec 1982.

N. Sooch, et al. "18-Bit Stereo D/A Converter with Integrated Digital and Analog Filters," *91st Audio Engineering Society Conv.,* Reprint # 3113(y-1), October 1991.

K. R. Stafford, P. R. Gray, and R. A. Blanchard. "A Complete Monolithic Sample/Hold Amplifier," *IEEE J. of Solid-State Circuits,* Vol. 9, no. 6, pp. 381–387, December 1974.

P. Vorenkamp and J. P. Verdaasdonk. "Fully bipolar, 120-MSample/s 10-b track-and-hold circuit," *IEEE J. Solid-State Circuits*, Vol. 27, pp. 988–992, Jul. 1992.

M. Wakayama, H. Tanimoto, T. Tasai, and Y. Yoshida. "A 1.2-μm BiCMOS Sample-and-Hold Circuit with a Constant-Impedance Slew-Enhanced Sampling Gate," *IEEE J. of Solid-State Circuits,* Vol. 27, no. 12, pp. 1697–1708, December 1992.

11.9 PROBLEMS

Unless otherwise stated, assume the following:

- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8\ V$
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}\,(\Omega) = 8{,}000\ L\ (\mu m)/I_D\ (mA)$ in active region,
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9\ V$
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}\,(\Omega) = 12{,}000\ L\ (\mu m)/I_D\ (mA)$ in active region,
  $C_j = 4.5 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.5 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$

11.9.1 Section 11.1: Performance of Sample-and-Hold Circuits

11.1 A S/H circuit has a droop rate of 15 V/s. Assuming the circuit spends 50% of its time in hold mode and a clock frequency of 10 MHz, how much droop is observed on each clock cycle?

11.2 A S/H circuit is tested at a clock frequency of 20 MHz with a 21-MHz input signal. At what frequency is the output?

11.9.2 Section 11.2: MOS Sample-and-Hold Basics and Section 11.3: Examples of CMOS S/H Circuits

11.3 Using the transient analysis of SPICE with a sinusoidal input signal of 2 V peak-to-peak, find the offset, gain, and linearity error of the simple S/H of Fig. 11.3.

11.4 For the T/H of Fig. 11.4, assume V_{in} is a 20-MHz sinusoid with a 2 $V_{p\text{-}p}$ amplitude. Also assume that ϕ_{clk} is a 100-MHz square wave having a peak amplitude of ±2.5 V with rise and fall times of 1.5 ns. What is the maximum time difference between the turn-off times of the n-channel and p-channel transistors? Ignore the body effect.

11.5 The T/H of Fig. 11.5 has transistors that are 10 μm/0.8 μm and a 1-pF hold capacitor. Assume the clock waveforms are fast enough that the channel charge of the transistors is evenly distributed between the two junctions. Compare the final hold pedestal between the case when the dummy switch turns on substantially before the sampling switch turns off, and the case when it turns on substantially after the sampling switch.

11.6 Assume the second stage of the S/H of Fig. 11.10 has a gain of –30 and a unity-gain frequency of 100 MHz. Also assume the input signal is a 5-MHz sinusoid with a 2 $V_{p\text{-}p}$ amplitude and the clock signals are at 20 MHz and go between +2.5 V and –2.5 V, with rise and fall times of 1.5 ns. What is the sampling time uncertainty?

11.7 Assume the S/H of Fig. 11.12 is being realized in a BiCMOS process. This means that diodes and BJT transistors are available. Design a clock-driver circuit so that the clock voltages go from two diode drops above the input voltage (in track mode) to two diode drops below the output voltage (in hold mode).

11.8 Assume the opamp in the S/H of Fig. 11.15 has a finite gain of A_0. Derive the output voltage in terms of V_{in} and A_0 during hold mode (i.e., when ϕ_2 is high).

11.9 Derive the output voltage of the S/H of Fig. 11.19 at the end of ϕ_2 in terms of the input voltage at the end of ϕ_1, the output voltage at the end of ϕ_1 from the previous period, and the capacitor ratio C_1/C_2. Take the z-transform of this difference equation, and substitute $z = e^{j\omega T}$ to find the frequency-domain transfer function of the S/H. Making the assumption that $e^{j\omega T} \equiv 1 + j\omega T$ for $\omega << (1/T) = f_{clk}$, where f_{clk} is the sampling frequency, show that

$$f_{-3\,dB} \cong \frac{1}{2\pi}\frac{C_1}{C_2} f_{clk}$$

11.9.3 Section 11.4: Bipolar and BiCMOS Sample-and-Holds

11.10 Assume the S/H of Fig. 11.21 has each of D_5 and D_6 replaced by two diodes in series. Show the voltages at all nodes for the cases of sampling a 1 V input and a –1 V input for before as well as after track mode.

11.11 In the S/H of Fig. 11.23, show that the hold-mode feedthrough from the base of Q_6 to the output is given by

$$A_{feedthrough} = \frac{C_{be6}}{C_{hld} + C_{be6}}\left(1 - \frac{C_F}{C_{be6}}\right)$$

and that, therefore, feedthrough is eliminated if $C_F = C_{be6}$.

11.9.4 Section 11.5: Translinear Gain Cell

11.12 Consider the gain cell shown in Fig. 11.24, where I_2 is fixed to a bias current of 200 μA while I_1 varies such that the gain changes from 0.5 to 4. What are the extreme values of I_1 needed? Also, what is the maximum peak current that should be used for i_{o1}, assuming all transistors should remain in the active region with at least 20 μA of emitter current?

11.9.5 Section 11.6: Translinear Multiplier

11.13 Consider the translinear multiplier shown in Fig. 11.25, where $I_1 = 100\ \mu A$, $I_2 = 200\ \mu A$, and i_1 is set such that the gain $i_o/i_2 = 0.5$. If a common-mode noise signal of $5\ \mu A$ appears on I_1, what is the size of the output signal due to that noise signal? Repeat for a common-mode noise signal on I_2.

11.14 Repeat Problem 11.13 when the β's of the transistors' worst-case mismatch is 5 percent.

11.15 Design a four-quadrant translinear multiplier having differential voltage-current conversion stages based on emitter degeneration at both inputs and a differential voltage output. All transistors should be biased at 0.2 mA. Design the multiplier so that when both differential inputs are 1 V the differential output should be 1 V. Assume a 5-V power supply is available.

CHAPTER

12 Continuous-Time Filters

Integrated analog filters can be realized in either discrete time or continuous time. Discrete-time filters and signal processing are addressed in Chapter 13 and their implementation, switched-capacitor circuits, are discussed in Chapter 14. In switched-capacitor circuits, although the signals remain continuous in voltage (i.e., they are never quantized) they require sampling in the time domain. Because of this time-domain sampling, the clock rate must always be at least twice that of the highest frequency being processed to eliminate aliasing. In fact, typically the clock rate is much greater than twice the signal bandwidth, to reduce the requirements of an anti-aliasing filter. As a result, switched-capacitor filters are limited in their ability to process high-frequency signals.

This chapter focuses on *continuous-time filtering.* As their name suggests, continuous-time filters have signals that remain continuous in time and that have analog signal levels. Since no sampling is required, continuous-time filters have a significant speed advantage over their switched-capacitor counterparts. However, continuous-time filters do have some disadvantages. One is their need for tuning circuitry. Tuning is required because their filter coefficients are determined by the product of two dissimilar elements, such as capacitance and resistance (or transconductance) values. Thus, whereas switched-capacitor filter coefficients are determined by capacitance ratios and therefore have accuracies of 0.1 percent, integrated continuous-time coefficients are initially set to only about 30-percent accuracy! Fortunately, tuning circuitry can set the filter coefficient accuracy to around 1 percent at the cost of more (and sometimes complex) circuitry, as we will see in Section 12.9. Another practical disadvantage exhibited by most continuous-time filters is their relatively poor linearity and noise performance. Although low-frequency switched-capacitor filters have been realized with nonlinear distortion and noise performance better than 90 dB, high-frequency continuous-time filters typically have distortion plus noise performance not much better than 60 dB (often it is much worse). However, there are numerous high-speed applications in which distortion and noise performances are not too demanding, such as many data communication and video circuits.

> ***Key Point:*** *Continuous-time integrated analog filters process signals that are both continuous in amplitude and time. Discrete-time analog filters process signals sampled in time, and therefore generally operate on lower-frequency signals than their continuous-time counterparts.*

12.1 INTRODUCTION TO CONTINUOUS-TIME FILTERS

The basics of first- and second-order transfer functions were presented in Section 4.1. In analog filters, these transfer functions are realized by combinations of integrators, summers, and gain elements. These are the fundamental building blocks of integrated continuous-time filters. Combining these elements, any rational transfer function may be realized.

> ***Key Point:*** *Integrators, summers, and gain stages are the fundamental building blocks of analog filters.*

Key Point: *Any rational transfer function with real-valued coefficients may be factored into first- and second-order terms and, therefore, implemented as a cascade of first- and second-order filters.*

Any rational transfer function with real-valued coefficients may be factored into first- and second-order terms that also have real-valued coefficients, so any order filter can be realized by cascading first- and second-order sections. Such a cascade is common in practice, and we therefore restrict ourselves to a discussion of first- and second-order filters here.

12.1.1 First-Order Filters

Key Point: *One integrator is required for each pole in an analog filter.*

A block diagram for a general first-order continuous-time filter is shown in Fig. 12.1. It is composed of one integrator, one summer, and up to three gain elements. In general, the number of integrators required is equal to the order of the desired transfer function.

It can be shown that the transfer function for this block diagram is given by

$$H(s) \equiv \frac{V_{out}(s)}{V_{in}(s)} = \frac{k_1 s + k_0}{s + \omega_0} \tag{12.1}$$

It has a real pole at a frequency ω_0. To ensure the pole is in the left half of the s-plane, the value of ω_0 must be positive. The dc gain is $H(0) = k_0/\omega_0$, and the gain at very high frequencies is $H(\infty) = k_1$.

12.1.2 Second-Order Filters

Second-order filters are also called biquadratic filters or biquads for short because in their most general form their transfer function is the ratio of two quadratic polynomials. The block diagram of a practical implementation is shown in Fig. 12.2. The transfer function for this second-order filter is given by

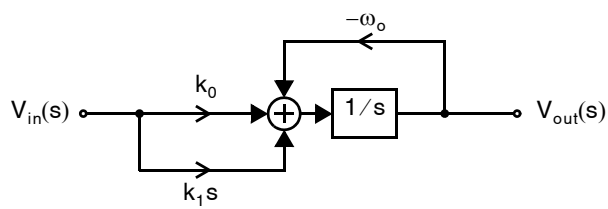


Fig. 12.1 A block diagram of a general first-order continuous-time filter.

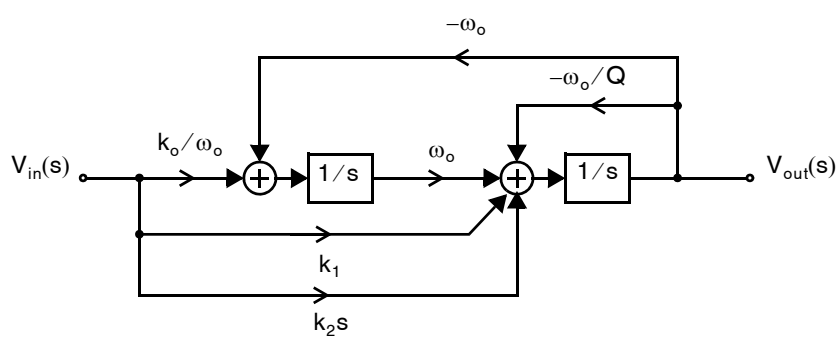


Fig. 12.2 A block diagram of a general second-order (biquadratic) continuous-time filter.

$$H(s) \equiv \frac{V_{out}(s)}{V_{in}(s)} = \frac{k_2 s^2 + k_1 s + k_o}{s^2 + \left(\frac{\omega_o}{Q}\right)s + \omega_o^2} \tag{12.2}$$

The system's poles, like any linear system, can be determined by considering the system's undriven behavior, $V_{in} = 0$, and are therefore independent of the feed-in gains, $k_{1,2,3}$. Two integrators are needed to realize the two poles. To ensure stability, ω_0/Q must be positive, which implies that one of the two integrators in Fig. 12.2 must have negative feedback around it, making it lossy. A large feedback coefficient ω_0/Q implies a very lossy integrator and, therefore, a low Q-factor. With $Q < 1/2$ the two poles are real, whereas with $Q > 1/2$ the poles form a complex-conjugate pair arranged as illustrated in Fig. 4.9.

Key Point: *A popular biquad implementation is two integrators in a loop, one being lossy. The gains around the loop determine its resonant frequency, the integrator loss determines the Q-factor, and the input gains and locations determine the transfer function zeros and gain.*

The numerator of (12.2) and hence the zeros of the biquad are determined by the feed-in gains $k_{1,2,3}$. By omitting some of the feed-in branches of Fig. 12.2, it can be used to realize either low-pass, high-pass, or band-pass biquads.

EXAMPLE 12.1

Find the gain coefficients in the biquad of Fig. 12.2 required to realize a low-pass transfer function with a dc gain of 6 dB, $Q = 1/\sqrt{2}$, and a 3-dB bandwidth of $\omega_{3\text{-dB}} = 10^6$ rad/sec.

Solution

A low-pass biquad is obtained with $k_1 = k_2 = 0$.

$$H(s) = \frac{k_o}{s^2 + \left(\frac{\omega_o}{Q}\right)s + \omega_o^2} \tag{12.3}$$

The 3-dB bandwidth is the frequency at which the magnitude of the denominator is $\sqrt{2}$ times larger than its dc value, which for $Q = 1/\sqrt{2}$ is the solution to

$$\begin{aligned} \left| -\omega_{3\text{-dB}}^2 + j\sqrt{2}\omega_0\omega_{3\text{-dB}} + \omega_o^2 \right| &= \sqrt{2}\omega_o^2 \\ \Rightarrow \omega_{3\text{-dB}} &= \omega_0 \end{aligned} \tag{12.4}$$

Hence, in this case $\omega_0 = 10^6$. Finally, for a dc gain of 6 dB, $k_0/\omega_0^2 = 2 \Rightarrow k_0 = 2\omega_0^2$. The block diagram of Fig. 12.2 may be simplified somewhat by incorporating a gain ω_0 into each integrator, resulting in the biquad of Fig. 12.3.

12.2 INTRODUCTION TO G_m-C FILTERS

This section gives a brief introduction to G_m-C filters. We will see here that the main circuit building blocks of G_m-C filters are transconductors. The basic functions of transconductors are described in this section, and circuit techniques for realizing transconductors are presented in the following sections.

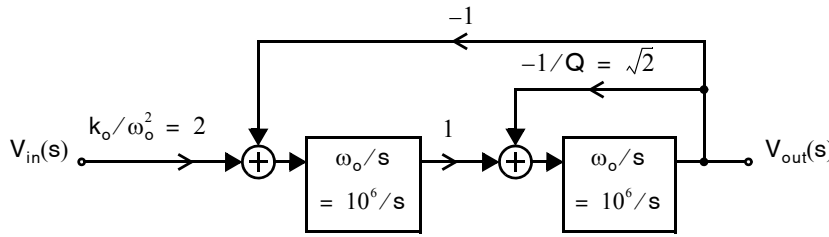


Fig. 12.3 The biquad filter of Example 12.1.

12.2.1 Integrators and Summers

Key Point: *In G_m-C filters, integrators are realized with a transconductor, which produces a output current proportional to the input voltage, and a capacitor, which integrates the current to produce the output voltage.*

An integrator is the main building block for most continuous-time filters. To realize an integrator in G_m-C technology, a transconductor and a capacitor can be used as shown in Fig. 12.4. A transconductor is essentially a transconductance cell (an input voltage creates an output current) with the additional requirement that the output current be *linearly related to the input voltage* . Here, the output of the transconductor is the current i_o, and, in the ideal case, both the input and output impedance of this block are infinite. The output current is linearly related to the differential input voltage signal by

$$i_o = G_m v_i \tag{12.5}$$

where G_m is the transconductance of the cell. It should be re-emphasized that this relationship is different from that of an operational-transconductance amplifier (OTA) in that here, the output current should be *linearly* related to the input voltage. In addition, a transconductor should have a well-known (or at least controllable) transconductance value. In the case of an OTA, the transconductance does not necessarily need to be a linear function of the input voltage or a well-controlled value (though usually the transconductance value should be as large as possible).

This output current is applied to the integrating capacitor, C_1, resulting in the voltage across C_1 being given by

$$V_o = \frac{I_o}{sC_1} = \frac{G_m V_i}{sC_1} \tag{12.6}$$

Defining ω_{ti} to be the unity-gain frequency of the integrator, we can write

$$V_o \equiv \left(\frac{\omega_{ti}}{s}\right) V_i = \left(\frac{G_m}{sC_1}\right) V_i \tag{12.7}$$

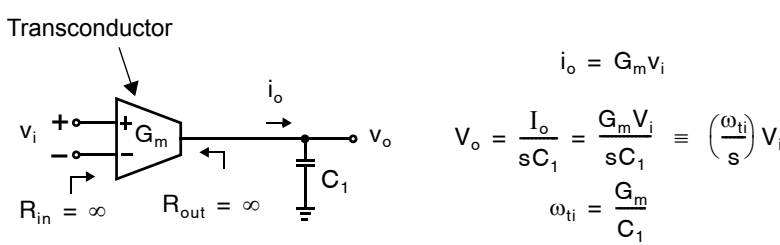


Fig. 12.4 A single-ended G_m-C integrator.

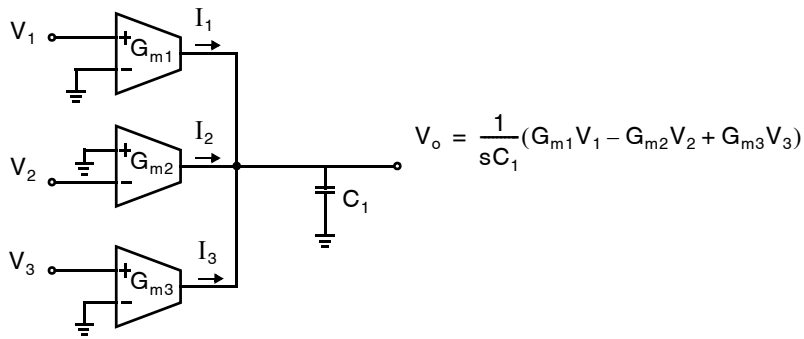


Fig. 12.5 A three-input, single-ended integrator/summer circuit.

Thus, we see that the output voltage is equal to the integration of the differential input voltage multiplied by the integrator unity-gain frequency, which is given by

$$\omega_{ti} = \frac{G_m}{C_1} \tag{12.8}$$

To realize a more general integrator/summer circuit, one need only connect multiple transconductors in parallel since the output currents all sum together. An example of a three-input integrator/summer circuit is shown in Fig. 12.5.

Key Point: *Current signals are readily summed by simply shorting them together.*

EXAMPLE 12.2

What transconductance is needed for an integrator that has a unity-gain frequency of 20 MHz, assuming a 2-pF integrating capacitor?

Solution

Rearranging (12.8), we have

$$\begin{aligned} G_m &= 2\pi \times 20 \text{ MHz} \times 2 \text{ pF} \\ &= 0.251 \text{ mA/V} \end{aligned} \tag{12.9}$$

Note that this transconductance value corresponds to

$$G_m = 1/3.98 \text{ k}\Omega \tag{12.10}$$

which, not surprisingly, is related to the unity-gain frequency by

$$2\pi \times 20 \text{ MHz} = \frac{1}{3.98 \text{ k}\Omega \times 2 \text{ pF}} \tag{12.11}$$

In other words, the required transconductance value is the inverse of the resistance value needed to form an RC time constant that corresponds to the desired unity-gain frequency.

12.2.2 Fully Differential Integrators

Key Point: *Fully-differential circuits are often preferred in analog integrated circuits as they provide improved linearity and noise immunity.*

Although the preceding blocks have all been shown with single-ended signals,[1] in most integrated applications it is desirable to keep the signals fully differential. As discussed in Section 6.7, fully differential circuits have better noise immunity and distortion properties. Thus, fully differential circuits are the primary focus of the remainder of this chapter.

A fully differential transconductor has two outputs—one positive output (current flowing out for a positive input voltage) and one negative output (current flowing in for a positive input voltage). With this new degree of freedom (having two outputs instead of just one), a fully differential integrator can be realized in one of two ways, as shown in Fig. 12.6.

Note that the integrator shown in Fig. 12.6(*a*) requires one-fourth the capacitance needed for the integrator of Fig. 12.6(*b*) to achieve the same integrator coefficient. Although at first glance, this increase in size appears to be a disadvantage of the circuit shown in Fig. 12.6(*b*), it should be mentioned that fully differential circuits require some method of common-mode feedback. This requirement is due to the fact that, although the differential signal remains stable due to connections of the filter feedback network, the common-mode signal is free to drift if it is not also controlled using some feedback circuit. As a result, one must also be concerned with the stability of the common-mode feedback circuit. Since dominant pole compensation using capacitors on the output nodes is often used to stabilize the common-mode network, the integrator shown in Fig. 12.6(*b*) has the advantage that the integrating capacitors, $2C_1$, can also be used for this purpose. Also, recall that the single capacitor of the integrator shown in Fig. 12.6(*a*) must be realized using integrated capacitors, and thus top- and bottom-plate parasitic capacitances exist. Since the bottom-plate parasitic capacitance can be quite large (up to 20 percent of the capacitance value), to maintain symmetry, C_1 is often realized using two parallel capacitors, as shown in Fig. 12.7. We take the parallel combination of the two capacitances,

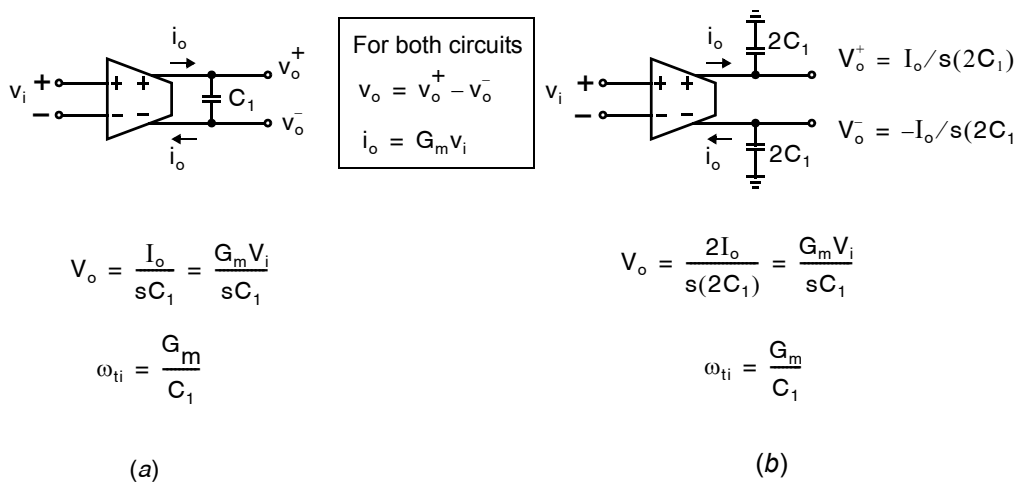
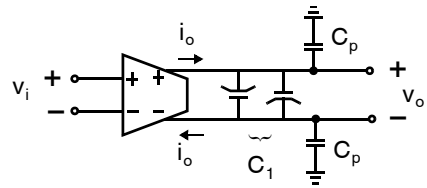


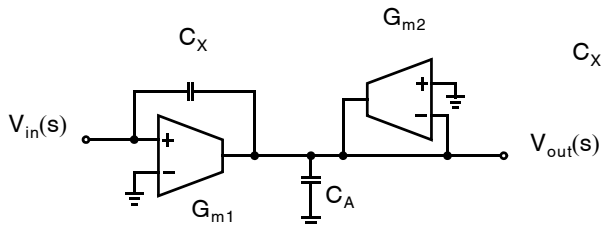
Fig. 12.6 Fully differential G_m-C integrators: (*a*) single capacitor, (*b*) two capacitors.

1. The transconductor input may be differential, but its output is single ended, and single-ended signals are typically applied to the transconductor inputs.



$$V_o = \frac{G_m V_i}{s(C_1 + C_p/2)} \qquad \omega_{ti} = \frac{G_m}{(C_1 + C_p/2)}$$

Fig. 12.7 A fully differential integrator maintaining symmetry and showing parasitic capacitances.



$$C_X = \left(\frac{k_1}{1 - k_1}\right) C_A \quad \text{where } 0 \le k_1 < 1$$

$$G_{m1} = k_0(C_A + C_X)$$

$$G_{m2} = \omega_o(C_A + C_X)$$

Fig. 12.8 A single-ended first-order G_m-C filter.

which equals C_1, and include the effect of the parasitic capacitances, C_p, to obtain the output voltage, given by

$$V_o = \frac{G_m V_i}{s(C_1 + C_p/2)} \tag{12.12}$$

Thus, we see here that the parasitic capacitances affect the integration coefficient and must be taken into account.

12.2.3 First-Order Filter

A single-ended G_m-C filter that realizes the first-order filter depicted in Fig. 12.1 is shown in Fig. 12.8. Note that the gains in Fig. 12.1 are provided by the transconductors, the integration is performed by the capacitor C_A, and the capacitor C_x corresponds to the differentiating feed-in path k_1 in Fig. 12.1.

Key Point: *Differentiating signal paths are sometimes provided by capacitive feed-ins.*

The transfer function for this first-order G_m-C filter is found by writing a current equation for the output node, $V_{out}(s)$:

$$G_{m1}V_{in}(s) + sC_X[V_{in}(s) - V_{out}(s)] - sC_A V_{out}(s) - G_{m2}V_{out}(s) = 0 \tag{12.13}$$

Rearranging to find $V_{out}(s)/V_{in}(s)$, we have

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{sC_X + G_{m1}}{s(C_A + C_X) + G_{m2}} = \frac{\left[s\left(\frac{C_X}{C_A + C_X}\right) + \left(\frac{G_{m1}}{C_A + C_X}\right)\right]}{\left[s + \left(\frac{G_{m2}}{C_A + C_X}\right)\right]} \tag{12.14}$$

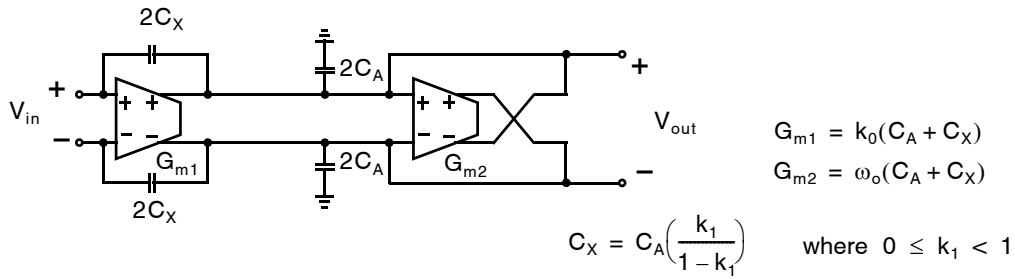


Fig. 12.9 A fully differential general first-order G_m-C filter.

Equating (12.14) with (12.1) results in the following relationships:

$$C_X = \left(\frac{k_1}{1 - k_1}\right) C_A \tag{12.15}$$

$$G_{m1} = k_0(C_A + C_X) \tag{12.16}$$

$$G_{m2} = \omega_o(C_A + C_X) \tag{12.17}$$

It should be noted from (12.15) that, for C_X to remain a positive value, the term k_1 must satisfy $0 \leq k_1 < 1$. What this restriction implies is that the shown circuit cannot realize a filter that has a high-frequency gain that is either negative or greater than 1. Such a result should come as no surprise since the high-frequency gain of this circuit is simply a capacitor-divider circuit, and hence the restriction on k_1.

Finally, the equivalent fully differential first-order G_m-C filter is shown in Fig. 12.9. Note that capacitor sizes are doubled here, in relation to the single-ended case, so that the same design equations occur. Also note that the same restriction holds on k_1, although a negative value of k_1 can be realized by cross-coupling the feed-in capacitor pair, C_X.

EXAMPLE 12.3

Based on Fig. 12.9, find component values for a first-order filter with a dc gain of 0.5, a pole at 20 MHz, and a zero at 40 MHz. Assume the integrating capacitors are sized $C_A = 2$ pF.

Solution

A first-order filter with a zero at 40 MHz and a pole at 20 MHz is given by

$$H(s) = \frac{k(s + 2\pi \times 40 \text{ MHz})}{(s + 2\pi \times 20 \text{ MHz})} \tag{12.18}$$

The leading coefficient is determined by setting the dc gain to 0.5, resulting in $k = 0.25$. Thus, we have

$$H(s) = \frac{0.25s + 2\pi \times 10 \text{ MHz}}{s + 2\pi \times 20 \text{ MHz}} \tag{12.19}$$

and equating this transfer function to (12.1) results in

$$k_1 = 0.25 \tag{12.20}$$

$$k_0 = 2\pi \times 10^7 \tag{12.21}$$

$$\omega_o = 4\pi \times 10^7 \tag{12.22}$$

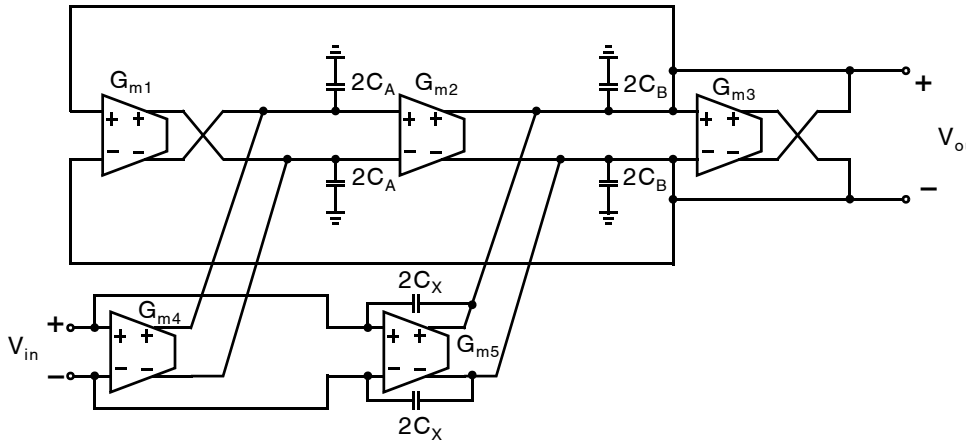


Fig. 12.10 A general second-order filter using fully differential G_m-C techniques.

Using the design equations shown in Fig. 12.9, we find

$$C_X = 2\text{ pF} \times \frac{0.25}{1 - 0.25} = 0.667\text{ pF} \tag{12.23}$$

$$G_{m1} = 2\pi \times 10^7 \times 2.667\text{ pF} = 0.168\text{ mA/V} \tag{12.24}$$

$$G_{m2} = 4\pi \times 10^7 \times 2.667\text{ pF} = 0.335\text{ mA/V} \tag{12.25}$$

where $C_A = 2$ pF, as given.

12.2.4 Biquad Filter

A fully-differential biquadratic G_m-C filter is shown in Fig. 12.10. The transfer function of this biquad filter is found to be given by

$$H(s) \equiv \frac{V_{out}(s)}{V_{in}(s)} = \frac{s^2\left(\frac{C_X}{C_X + C_B}\right) + s\left(\frac{G_{m5}}{C_X + C_B}\right) + \left(\frac{G_{m2}G_{m4}}{C_A(C_X + C_B)}\right)}{s^2 + s\left(\frac{G_{m3}}{C_X + C_b}\right) + \left(\frac{G_{m1}G_{m2}}{C_A(C_X + C_B)}\right)} \tag{12.26}$$

Equating the coefficients of (12.2) and (12.26) results in

$$k_2 = \frac{C_X}{C_X + C_B} \tag{12.27}$$

$$k_1 = \frac{G_{m5}}{C_X + C_B} \tag{12.28}$$

$$k_0 = \frac{G_{m2}G_{m4}}{C_A(C_X + C_B)} \tag{12.29}$$

and

$$\omega_o^2 = \frac{G_{m1}G_{m2}}{C_A(C_X + C_B)} \tag{12.30}$$

For Q, we note that

$$\frac{\omega_o}{Q} = \frac{G_{m3}}{C_X + C_b} \tag{12.31}$$

and so, by using (12.30), we can solve for Q, resulting in

$$Q = \sqrt{\left(\frac{G_{m1}G_{m2}}{G_{m3}^2}\right)\left(\frac{C_X + C_B}{C_A}\right)} \tag{12.32}$$

Although the preceding equations are for analyzing a given circuit, they can also be used for design, resulting in the following design equations:

$$C_X = C_B\left(\frac{k_2}{1 - k_2}\right) \qquad \text{where } 0 \le k_2 < 1$$

$$G_{m1} = \omega_o C_A \qquad G_{m2} = \omega_o(C_B + C_X) \qquad G_{m3} = \frac{\omega_o(C_B + C_X)}{Q}$$

$$G_{m4} = (k_o C_A)/\omega_o \qquad G_{m5} = k_1(C_B + C_X)$$

Note that for this design, there is a restriction on the high-frequency gain coefficient, k_2, similar to that occurring in the first-order case.

EXAMPLE 12.4

Find the transconductance and capacitor values for a second-order filter that has a bandpass response with a center frequency of 20 MHz, a Q value of 5, and a center frequency gain of 1. Assume that $C_A = C_B = 2$ pF.

Solution

A bandpass transfer function has the form

$$H(s) \equiv \frac{V_{out}(s)}{V_{in}(s)} = \frac{Gs\frac{\omega_o}{Q}}{s^2 + s\frac{\omega_o}{Q} + \omega_o^2} \tag{12.33}$$

where $G = 1$ is the gain at the center frequency. Using the facts that $\omega_o = 2\pi \times 20$ MHz and $Q = 5$, we find

$$k_1 = G\frac{\omega_o}{Q} = 2.513 \times 10^7 \text{ rad/s} \tag{12.34}$$

Since k_0 and k_2 are zero, we have $C_x = G_{m4} = 0$ and

$$G_{m1} = \omega_o C_A = 0.2513 \text{ mA/V} \tag{12.35}$$

$$G_{m2} = \omega_o(C_B + C_X) = 0.2513 \text{ mA/V} \tag{12.36}$$

and, in this case,

$$G_{m3} = G_{m5} = k_1 C_B = 50.27\ \mu A/V \tag{12.37}$$

12.3 TRANSCONDUCTORS USING FIXED RESISTORS

Two similar approaches for using resistors to linearize the relationship between differential input voltage and output current of a differential pair of transistors are shown in Fig. 12.11. To understand the basic operation of the two circuits, we first make the simplifying assumption that the V_{gs} voltages of the transistors are constant. As a result, we see that the differential input voltage, v_i, appears across the two $R_S/2$ resistors in Fig. 12.11(*a*) and across R_S in Fig. 12.11(*b*). Thus, the drain current I_{D1} is given by

$$I_{D1} = I_1 + \frac{v_i}{R_S} \tag{12.38}$$

and the drain current I_{D2} is given by

$$I_{D2} = I_1 - \frac{v_i}{R_S} \tag{12.39}$$

Then defining i_{o1} to be the difference in the drain currents from I_1 (as shown), we have

$$i_{o1} = \frac{v_i}{R_S} \tag{12.40}$$

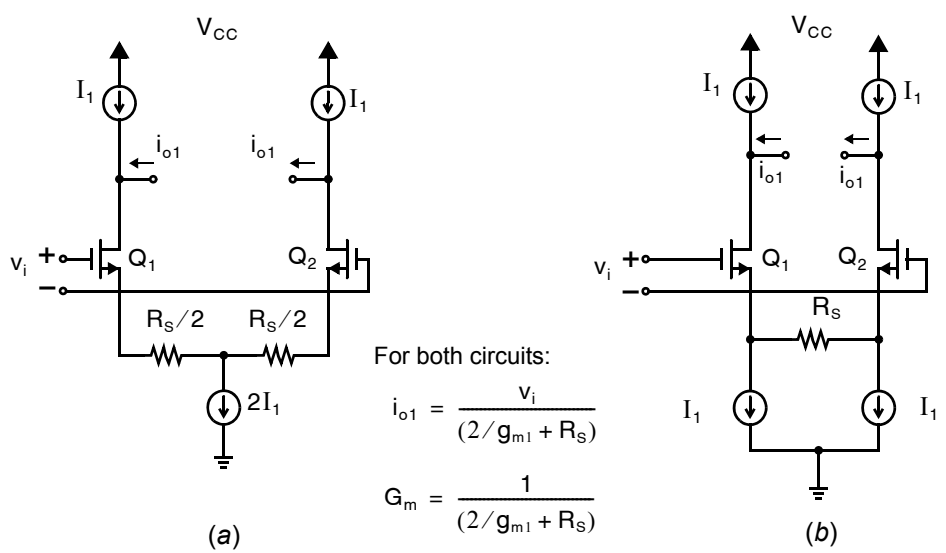


Fig. 12.11 Fixed transconductors using resistors. The circuit in (*a*) has smaller common-mode range and output swing than the circuit in (*b*) due to bias currents flowing through the $R_S/2$ resistors.

Key Point: *Source-degenerating a differential pair with a resistor linearizes the relationship between differential input voltage and output current. The result is a transconductor with a value approximately equal to the degeneration resistance.*

Thus, we see that the transconductance of these two differential input circuits is a fixed value of $1/R_S$. One important difference between the two fixed transconductors of Fig. 12.11 is that, in the case of $v_i = 0$, a bias current of I_1 flows through the two $R_S/2$ resistors in Fig. 12.11(*a*), whereas no bias current is flowing through R_S in Fig. 12.11(*b*). As a result, the common-mode voltage of Fig. 12.11(*b*) can be closer to ground. For example, if a bias current of $I_1 = 100\ \mu A$ is used together with $R_S = 20\ k\Omega$, then the voltage drop across each of the two $R_S/2$ resistors would result in 1 V less of common-mode range. It should be noted here that the common-mode input voltage must be far enough above ground to account for the largest expected differential input voltage, v_i. Specifically, one must ensure that the circuit remains functional while the gate on one side of the differential pair is lower than the common-mode voltage by $v_i/2$.

A simple estimate of the maximum differential input range for both these circuits can be found by determining how much voltage exists across the linearizing resistors at the point where one transistor just turns off. For example, in Fig. 12.11(*a*), when Q_2 turns off due to a large positive input voltage, all of $2I_1$ flows through $R_s/2$ on the left and nothing flows through $R_s/2$ on the right. Thus, the voltage at the center of the two resistors equals that of the source of Q_2 (since no current flows right), and the maximum input voltage equals that voltage across the left $R_S/2$ resistor, giving

$$V_{i,max} = I_1 R_S \tag{12.41}$$

A similar argument gives the same maximum input voltage for the circuit of Fig. 12.11(*b*). It should be noted here that this simple estimate does not take into account that significant distortion occurs when this maximum input voltage is applied. Specifically, at $V_{i,max} = I_1 R_S$, one of the two transistors will be off, implying that V_{GS} has not remained a relatively constant value, as we assumed earlier in this section. Typically, the maximum input range that applies to the preceding fixed transconductors might be one-half that of (12.41), or, equivalently, $(I_1/2)R_S$.

So far, we have assumed that the transistor's V_{gs} voltages remain constant. Such an assumption is reasonably valid if the MOSFETs are replaced by bipolar transistors as shown in Fig. 12.12 due to the exponential current-voltage relationship of bipolar transistors. However, it is less accurate with a MOFET differential pair, whose

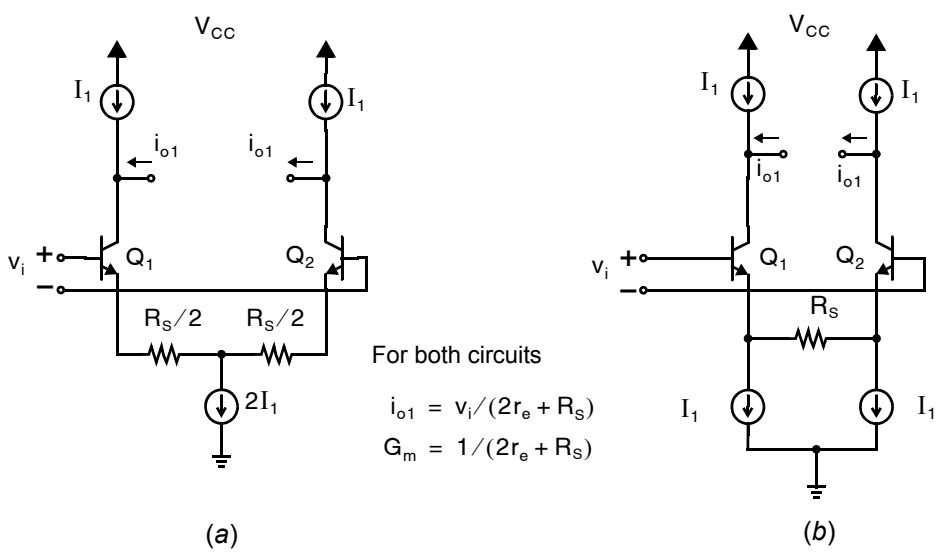


Fig. 12.12 Bipolar implementations of the transconductors shown in Fig. 12.11.

gate-source voltage changes appreciably with input voltage. In either the MOS or bipolar circuits, the assumption is reasonable when one restricts the minimum drain (or collector) current, $I_{D,min}$, in either transistor to be that which keeps the small-signal transconductance, g_m, large enough to ensure $1/g_m \ll R_S$, recalling that

$$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D} \gg 1/R_S \tag{12.42}$$

For the bipolar circuit, we have the simpler requirement,

$$r_{e,max} = \frac{V_T}{I_{E,min}} \ll R_S \tag{12.43}$$

or equivalently,

$$I_{E,min} \gg \frac{V_T}{R_S} \tag{12.44}$$

For example, if $R_S = 20$ kΩ and $V_T = 25$ mV, the minimum emitter current in either transistor should be much greater than 1.25 μA. A good question is, How much is "much greater than"? The answer to this question depends on how much distortion can be tolerated since, when $1/g_m$ becomes a comparable size to R_S, the varying V_{gs} or V_{be} voltage will result in the output current, i_o, being a significant nonlinear function of the input voltage, v_i. For example, if the maximum value of $1/g_m$ goes to 10 percent of that of R_S when the maximum applied input voltage is $(I_1/2)R_S$, then a linearity deviation of about 3 percent occurs. Thus, to attain a high linearity with the preceding circuits, the bias currents should be large and the input signal should be small (to ensure large values of g_m). Of course, the use of small input signals may result in noise becoming a problem.

In applications in which a moderate amount of nonlinearity can be tolerated, R_S might be chosen similar in size to the nominal bias value of $1/g_m$. In this case, the overall transconductance of this transconductor is most easily found using the hybrid T model for the input transistor pair. With such an approach, the output current is more accurately described by

$$i_{o1} = \frac{v_i}{1/g_{m1} + R_S + 1/g_{m2}} = \frac{v_i}{2/g_m + R_S} \tag{12.45}$$

Note that (12.45) reduces to (12.40) when R_S is chosen much larger than $2/g_m$. In summary, a more accurate formula for the transconductance of the circuits of Fig. 12.11 (and Fig. 12.12) is

$$G_m = \frac{i_{o1}}{v_i} = \frac{1}{2/g_m + R_S} \tag{12.46}$$

EXAMPLE 12.5

Consider the circuit of Fig. 12.12(*b*), where $R_S = 3$ kΩ (such that $G_m \approx 0.333$ mA/V) and the maximum differential input voltage is ±500 mV. Choose an appropriate bias current, I_1, such that $r_{e,max} < 0.1R_s$, and find a more precise value for G_m. Find the new value of I_1 and G_m if the constraint $r_{e,max} < 0.01R_s$ is used instead.

Solution

In the first case, we have $r_{e,max} < 0.1R_s = 300\ \Omega$. Thus, in the case of $v_i = 500$ mV, we require that $r_{e2} \approx 300\ \Omega$, which is attained for an emitter current of

$$I_{E2,min} = \frac{V_T}{r_{e,max}} = \frac{26\text{ mV}}{300\ \Omega} = 86.7\ \mu\text{A} \tag{12.47}$$

For this same input voltage, the resistor current is approximately

$$I_{RE} \cong \frac{500\ \text{mV}}{3\ \text{k}\Omega} = 167\ \mu\text{A} \tag{12.48}$$

and since the current through R_s plus I_{E2} must equal I_1 (according to the node currents at the emitter of Q_2), we have

$$I_1 = 167 + 86.7 = 253\ \mu\text{A} \tag{12.49}$$

In other words, a minimum bias current of $253\ \mu\text{A}$ should be chosen. For this value of bias current, the nominal value of r_e is given by

$$r_e = \frac{V_T}{I_1} = 103\ \Omega \tag{12.50}$$

resulting in

$$G_m = \frac{1}{(2r_e + R_s)} = 0.312\ \text{mA/V} \tag{12.51}$$

For the case $r_{e,max} < 0.01R_E$, one finds $I_1 = 1.03$ mA and $G_m = 0.328$ mA/V. Note here that the nominal value of r_e is now only $26\ \Omega$, which is significantly less than R_s. Although this second design will have less distortion than the first (about 10 times better), it uses much more power. Next, we look at using feedback to improve distortion without resorting to large bias currents.

One way to improve the distortion properties of the preceding fixed transconductors without using large bias currents and small input signal levels is to use an extra pair of amplifiers, as shown in Fig. 12.13. Here, the virtual ground in each opamp forces the source voltages of Q_1 and Q_2 to equal those of v_i^+ and v_i^-. As a result, the input voltage appears directly across R_S and does not depend on the V_{gs} voltages of Q_1 and Q_2. Thus, the transconductance of this circuit is given by

$$G_m = \frac{i_{o1}}{v_i^+ - v_i^-} = \frac{1}{R_S} \tag{12.52}$$

It should be mentioned here that the opamps can be quite simple (typically, a single stage) because they merely improve an input stage that is already reasonably linear.

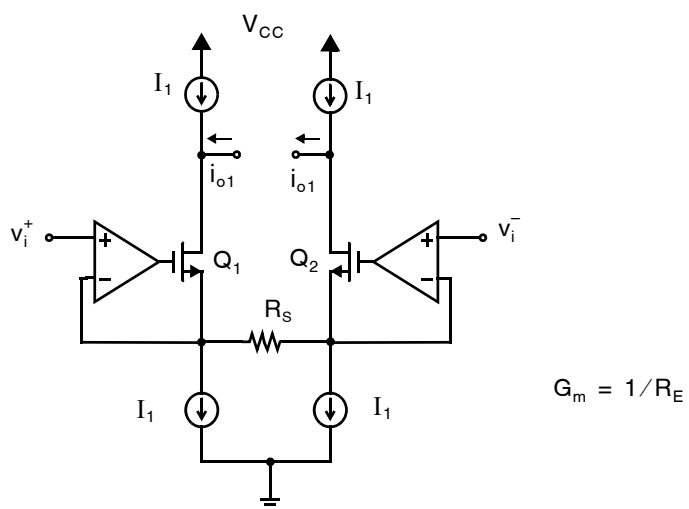


Fig. 12.13 Improving the linearity of a fixed transconductor through the use of opamps.

> ***Key Point:*** *Feedback may be used to more accurately reproduce the input voltage across the degeneration resistor or further improve the linearity of transconductors with fixed resistors.*

Another method for improving the linearity of a fixed transconductor is to force constant currents through Q_1 and Q_2, as shown in Fig. 12.14 for a bipolar input pair [Koyama, 1993]. Here, I_1 flows through each of Q_1 and Q_2 so that V_{be1} and V_{be2} are fixed, while Q_3 and Q_4 have varying collector currents corresponding to the current flow through R_s. The output currents can then be obtained at the collectors of Q_7 and Q_8, which mirror the current through Q_3 and Q_4. Thus, assuming the current mirror transistors consisting of Q_7 and Q_8 are equal in size to those of Q_3 and Q_4, the transconductance of this cell is given by

$$G_m = \frac{i_{o1}}{v_i^+ - v_i^-} = \frac{1}{R_s} \tag{12.53}$$

If the current mirror transistors are scaled in relation to Q_3 and Q_4, then the transconductance for this cell is also scaled by the same amount.

It should be noted that simple diode-connected transistors are not used at Q_3 and Q_4 since a feedback circuit is required to set the voltage level at the collectors of Q_1 and Q_2 without drawing much current. The feedback loop occurs as follows: If the voltage at the collector of Q_1 increases because I_{C1} is less than I_1, then V_{be3} also increases, forcing more current through Q_3 and Q_1 until I_{C1} once again equals I_1. Although the circuit shown results in each of Q_1 and Q_2 collector voltages being set to two diode drops above ground (i.e., $V_{be3} + V_{be5}$), a larger voltage can be obtained by adding diode-connected transistors that operate as voltage level-shifters. It is possible to combine parts of this fixed-transconductor circuit with a gain-cell stage, as was done in [Koyama, 1993]. The same linearization technique has been used in a CMOS variable transconductor that will be presented in Section 12.4.1, where the resistance R_S is realized using a MOS transistor operating in the triode region [Welland, 1994].

A problem with the transconductors shown so far is that there is no obvious mechanism for tuning the transconductance to ensure accuracy over process, voltage, and temperature variations. One solution is to provide an array of degeneration resistors R_S and somehow select the most appropriate value using switches. Alternatively, triode devices may be used in series with or in place of R_S, and their gate voltage used to tune the effective source degeneration. This is the topic of the following section.

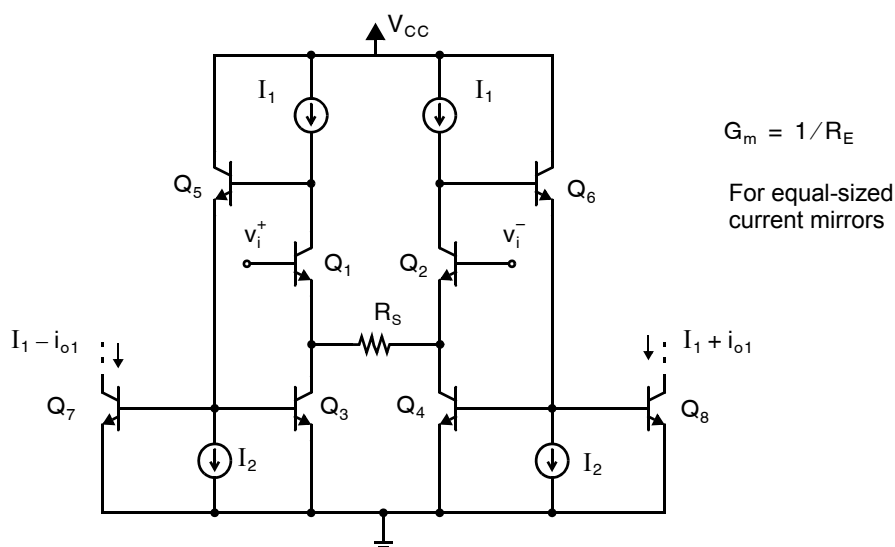


Fig. 12.14 Improving the linearity of a fixed transconductor by maintaining a constant V_{be} voltage for Q_1 and Q_2. Level shifts should be added at the emitters of Q_5, Q_6 if a larger input range is desired.

12.4 CMOS TRANSCONDUCTORS USING TRIODE TRANSISTORS

In this section, we discuss transconductor circuit techniques that rely on transistors operating in the triode region. It should be stated here that in the following circuits, not all transistors are in the triode region. As we will see, most of the transistors are biased in the active region, but the transconductance of the circuit relies on one or two key transistors biased in the triode region.

We begin this section by recalling the classical model equation for an n-channel transistor operating in the triode region.

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{12.54}$$

Also recall that a transistor remains in the triode region as long as the drain-source voltage is lower than the effective gate-source voltage. In other words, the following condition must be satisfied:

$$V_{DS} < V_{eff} \qquad \text{where } V_{eff} = V_{GS} - V_{tn} \tag{12.55}$$

We will see next that if the preceding equations were exact, then perfectly linear circuits could be realized. However, these equations are only mildly accurate for a modern short-channel process, and therefore some distortion occurs because higher-order terms are not modelled in the preceding equations. Hence, almost all practical continuous-time circuits use fully differential architectures to reduce even-order distortion products, leaving the third-order term to dominate.

12.4.1 Transconductors Using a Fixed-Bias Triode Transistor

We see from (12.54) that if a small drain-source voltage, V_{DS}, is used, the V_{DS}^2 term goes to zero quickly and the drain current is approximately linear with respect to an applied drain-source voltage. Thus, we can model a transistor in triode as a resistor given by

$$r_{DS} \equiv \left(\frac{\partial i_D}{\partial v_{DS}}\right)^{-1}\Bigg|_{v_{DS}=0} \tag{12.56}$$

which results in a small-signal resistance of

$$r_{DS} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)\left(V_{GS} - V_{tn}\right)} \tag{12.57}$$

Key Point: *Triode transistors can be used in placed of fixed resistors to make transconductors tunable, although their linearity is generally reduced.*

One approach for using this equivalent resistance to realize a transconductor with moderate linearity is shown in Fig. 12.15 [Welland, 1994]. This circuit has the same structure as that of the bipolar circuit of Fig. 12.14. However, here the bipolar transistors are replaced with their MOSFET counterparts (biased in the active region), and the resistor is replaced with a transistor operating in the triode region. Note that the circuit of Fig. 12.15 forces constant currents through Q_1 and Q_2 (as in the bipolar case) such that their individual gate-source voltages remain

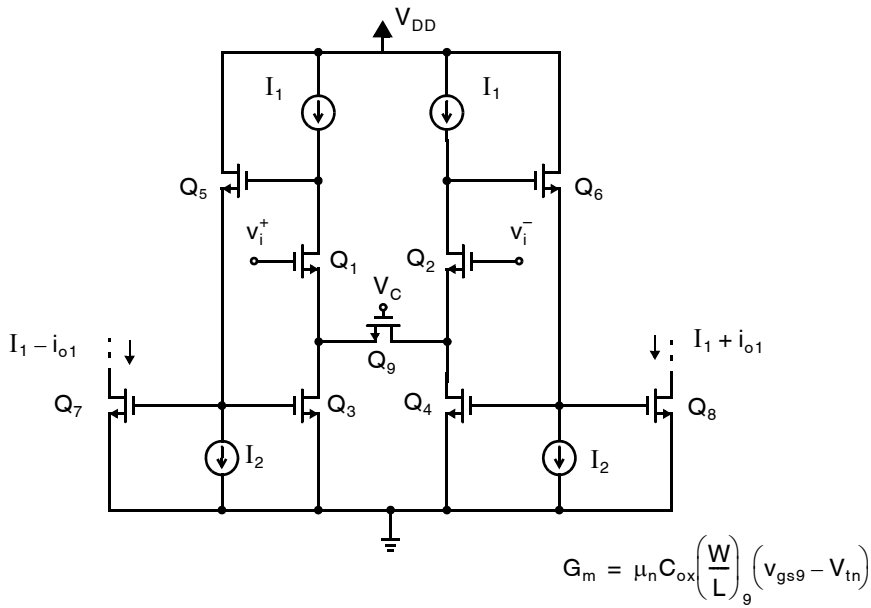


Fig. 12.15 A CMOS transconductor using Q_9 in the triode region.

constant. As a result, the transconductance of this cell is equal to the inverse of the small-signal resistance, r_{ds}, of Q_9. Using the small-signal resistance result of (12.57), the overall transconductance of this cell is given by

$$G_m = \frac{i_{o1}}{v_i^+ - v_i^-} = \mu_n C_{ox}\left(\frac{W}{L}\right)_9 (V_{gs9} - V_{tn}) \tag{12.58}$$

Thus, this transconductor has a variable transconductance value that can be adjusted by changing the value of V_{gs9}. As in the bipolar case, this transconductance value can be scaled by scaling the current mirrors Q_7 and Q_8 with respect to Q_3 and Q_4.

A similar approach but with reduced complexity and fewer p-channel inputs is shown in Fig. 12.16 [Kwan, 1991]. Here, a constant current of I_2 is forced through each of Q_1 and Q_2, whereas the dc bias current for each of Q_3 and Q_4 equals the difference between I_1 and I_2. The signal current, i_{o1}, also flows through Q_3 and Q_4, which is mirrored to the output transistors, Q_5 and Q_6. Note that this configuration has one less transistor stage around the input transistor loops. Specifically, the Q_5 and Q_6 source followers shown in Fig. 12.15 are not required in the circuit of Fig. 12.16, which usually results in a speed advantage for the latter.

An advantage of the transconductors in Figs. 12.15 and 12.16 (and other similar ones) is that they are easily modified to have multiple outputs by simply including additional output transistors. For example, the transconductor in Fig. 12.16 with two outputs is shown in Fig. 12.17. These additional transistors can be scaled for any desired ratio of output currents. These multiple outputs often allow filters to be realized using fewer transconductors. For example, a biquad similar to that in Fig. 12.10, but using multiple-output transconductors, is shown in Fig. 12.18. The transconductors are now associated with the outputs rather than the inputs. There is one transconductor per filter order plus an additional one for the input conversion from voltage to current. This approach has resulted in a savings of two transconductors. Not only does this approach save die area, but, often more importantly, it saves power.

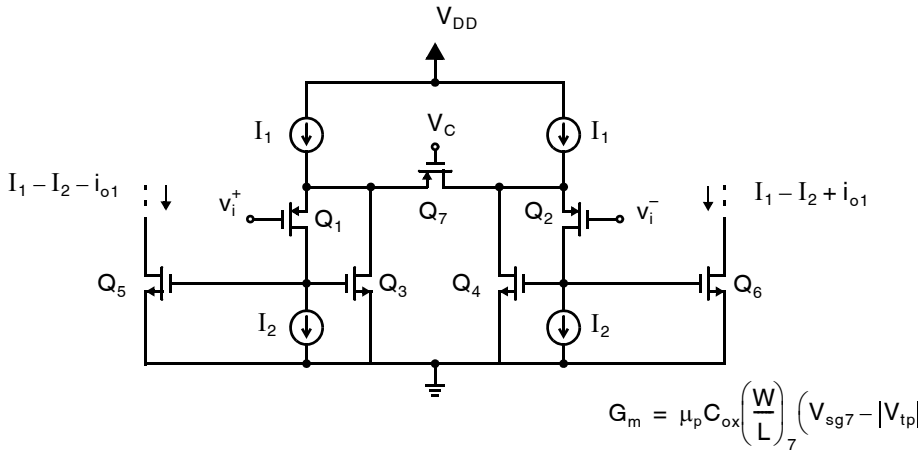


Fig. 12.16 A CMOS transconductor with p-channel inputs.

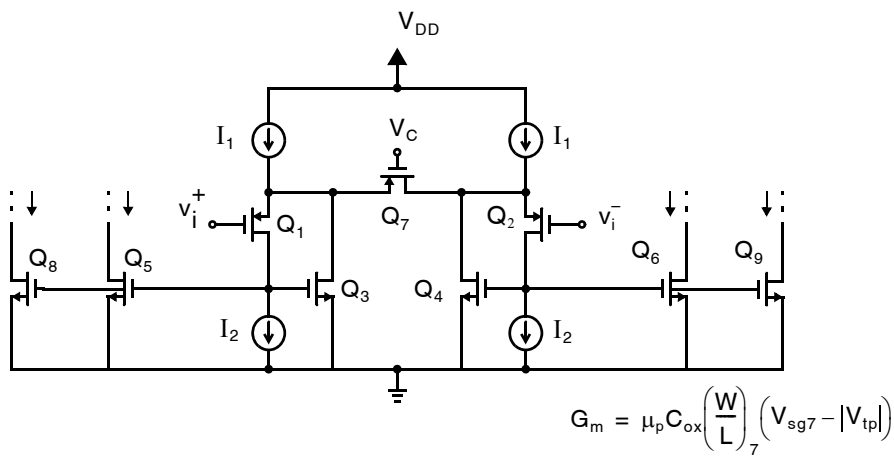


Fig. 12.17 A multiple-output CMOS transconductor.

12.4.2 Transconductors Using Varying Bias-Triode Transistors

Another way to linearize a MOSFET differential stage, using transistors primarily in the triode region, is shown in Fig. 12.19 [Krummenacher, 1988]. Note that here the gates of transistors Q_3, Q_4 are connected to the differential input voltage rather than to a bias voltage. As a result, triode transistors Q_3 and Q_4 undergo varying bias conditions to improve the linearity of this circuit (as discussed shortly). To see that Q_3 and Q_4 are generally in the triode region, we look at the case of equal input signals (i.e., $v_1 = v_2$), resulting in

$$v_x = v_y = v_1 - v_{GS1} = v_1 - (v_{eff1} + V_{tn}) \tag{12.59}$$

Thus, the drain-source voltages of Q_3, Q_4 are both zero. However, their gate-source voltages equal those of Q_1, Q_2, implying that they are indeed in the triode region.

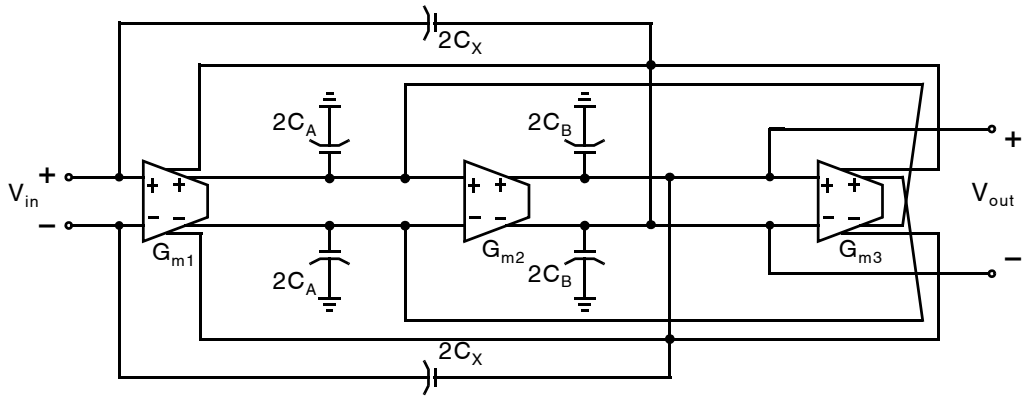
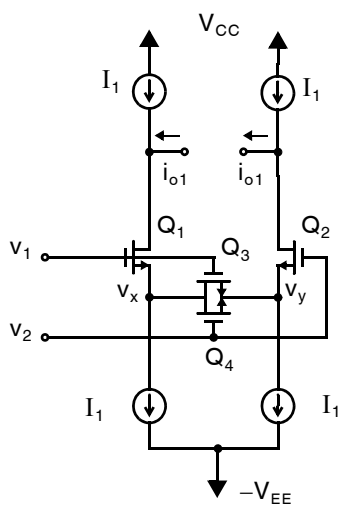


Fig. 12.18 A general biquad realized using multiple-output transconductors.



$$G_m = \frac{4k_1k_3\sqrt{I_1}}{(k_1 + 4k_3)\sqrt{k_1}}$$

Fig. 12.19 A linearized CMOS differential pair using transistors Q_3, Q_4 primarily in the triode region. The top pair of current sources can be set equal to the bottom pair through the use of a common-mode feedback.

To determine the small-signal transconductance of this stage, we note from (12.57) that the small-signal drain-source resistance of Q_3, Q_4 is given by

$$r_{ds3} = r_{ds4} = \frac{1}{2k_3(V_{GS1} - V_{tn})} \tag{12.60}$$

where we have assumed that Q_3 and Q_4 are matched devices and have defined

$$k_3 = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)_3 \tag{12.61}$$

Note that in this circuit we cannot ignore the effect of the varying gate-source voltages of Q_1 and Q_2 since their drain currents are not fixed to a constant value (as in Fig. 12.15). The small-signal source resistance values of Q_1 and Q_2 are given by

$$r_{s1} = r_{s2} = \frac{1}{g_{m1}} = \frac{1}{2k_1(V_{GS1} - V_{tn})} \tag{12.62}$$

Using the small-signal T model, we find the small-signal output current of i_{o1} to be equal to

$$i_{o1} = \frac{v_1 - v_2}{r_{s1} + r_{s2} + (r_{ds3} \parallel r_{ds4})} \tag{12.63}$$

Thus, defining $G_m \equiv i_{o1}/(v_1 - v_2)$, we have

$$G_m = \frac{1}{r_{s1} + r_{s2} + (r_{ds3} \parallel r_{ds4})} \tag{12.64}$$

which can be shown to be

$$G_m = \frac{4k_1k_3(V_{GS1} - V_{tn})}{k_1 + 4k_3} \tag{12.65}$$

Finally, to modify this result such that it is in terms of current, we note that

$$I_1 = k_1(V_{GS1} - V_{tn})^2 \tag{12.66}$$

or, equivalently,

$$(V_{GS1} - V_{tn}) = \sqrt{\frac{I_1}{k_1}} \tag{12.67}$$

which leads to

$$G_m = \frac{4k_1k_3\sqrt{I_1}}{(k_1 + 4k_3)\sqrt{k_1}} \tag{12.68}$$

Thus, we see that the transconductance of this transconductor can be tuned by changing the bias current, I_1. Note also that the transconductance is proportional to the square root of the bias current as opposed to a linear relation in the case of a bipolar transconductor.

To see how varying bias conditions for the triode transistors improves the linearity of this circuit, first consider the case of small differences in v_1 and v_2 (say, 100 mV). In this small-input-signal case, Q_3 and Q_4 essentially act as two source degeneration resistors. Thus, linearity improves (assuming $r_{ds3} >> r_{s1}$) over that of a simple differential pair, since transistors in triode are more linear than the source resistances of transistors in the active region. In other words, for a small-input-signal value, the linearity of this circuit is similar to that for a circuit in which the gates of Q_3, Q_4 are connected to fixed-bias voltages. However, if the gates of Q_3, Q_4 *were* connected to fixed-bias voltages, (12.54) shows that, as their drain-source voltage is increased, the drain current is reduced below that which would occur in a linear relationship due to the squared v_{DS} term. In addition, the changing values of V_{GS1} and V_{GS2} (due to current changes) also reduce the value of V_{DS}, and hence the transconductance of the cell for large input signals. To alleviate this reduction in transconductance for large input signals, the gates of Q_3 or Q_4 are connected to the input signals as shown. Now, as the input signal is increased, the small-signal resistance of one of the two triode transistors in parallel, Q_3 or Q_4, is reduced. This reduced resistance attempts to increase the transconductance, resulting in a partial cancelling of the decreasing transconductance value. Thus, if a proper ratio of k_1/k_3 is chosen, a more stable transconductance (and hence better linearity) is achieved. For a 4-MHz filter presented in [Krummenacher, 1988], a ratio of $k_1/k_3 = 6.7$ is used to obtain distortion levels around 50 dB below the maximum output signal level of 350

mVrms. Also, an extension of this approach is presented in [Silva-Martinez, 1991], where the input signal is split over two input stages to further improve the distortion performance.

EXAMPLE 12.6

Consider the circuit of Fig. 12.19 with the following values: $I_1 = 100\ \mu A$, $\mu_n C_{ox} = 96\ \mu A/V^2$, $(W/L)_1 = (W/L)_2 = 20$, $(W/L)_3 = (W/L)_4 = 3$, and $v_2 = 0\ V$.

1. Assuming a perfectly linear transconductor, use (12.68) to find i_{o1} when v_1 equals 2.5 mV and 250 mV.
2. Assume the gates of Q_3, Q_4 are connected to ground and use classical models for both the triode and active regions. Find the true value of i_{o1} when v_1 equals 2.5 mV and 250 mV. Compare your results with those in part 1.
3. Repeat part 2, assuming the gates of Q_3, Q_4 are connected to the input signals as shown in Fig. 12.19.

Solution

1. For the given values, we have

$$k_1 = \frac{96}{2} \times 20 = 960\ \mu A/V^2 \tag{12.69}$$

$$k_3 = \frac{96}{2} \times 3 = 144\ \mu A/V^2 \tag{12.70}$$

and substituting the above values in (12.68), we have

$$G_m = \frac{4k_1k_3\sqrt{100}}{(k_1 + 4k_3)\sqrt{k_1}} = 116.2\ \mu A/V \tag{12.71}$$

Now, to find i_{o1}, we use the definition of G_m to find

$$i_{o1} = G_m(v_1 - v_2) = G_m v_1 \tag{12.72}$$

Thus, for $v_1 = 2.5$ mV, we have $i_{o1} = 0.2905\ \mu A$. However, if $v_1 = 250$ mV, then $i_{o1} = 29.05\ \mu A$. In other words, for a 250-mV input voltage, the total drain current of Q_1 would be 129.05 μA, whereas that for Q_2 would be 70.95 μA.

2. To find i_{o1} in this case, we first note that the sum of the drain currents in Q_1 and Q_2 must equal the total bias current, 200 μA. In other words,

$$i_{D1} + i_{D2} = 2I_1 = 200\ \mu A \tag{12.73}$$

And since (using $v_{effi} = v_{GSi} - V_{tn}$) the drain currents in Q_1 and Q_2 are given by

$$i_{D1} = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)_1 (v_{eff1}^2) \quad \text{and} \quad i_{D2} = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)_2 (v_{eff2}^2) \tag{12.74}$$

we can write (12.73) as

$$v_{eff1}^2 + v_{eff2}^2 = \frac{5}{24}\ V^2 \tag{12.75}$$

The current from node v_x to v_y through the two triode transistors is given by

$$i_{xy} = 2 \times i_{D3} = 2\mu_n C_{ox}\left(\frac{W}{L}\right)_3 \left(v_{eff2}v_{xy} - \frac{v_{xy}^2}{2}\right) \tag{12.76}$$

where we have noted that $v_{eff3} = v_{eff2}$. The sum of i_{xy} and i_{D2} must equal I_1, resulting in

$$960v_{eff2}^2 + 576\left(v_{eff2}v_{xy} - \frac{v_{xy}^2}{2}\right) = 100 \tag{12.77}$$

Finally, we note that the input signal is equal to the following sum (the threshold voltage terms cancel):

$$-v_{eff2} + v_{xy} + v_{eff1} = v_{in} \tag{12.78}$$

Equations (12.75), (12.77), and (12.78) can be solved using an equation solver and then choosing the solution that has all positive values.

For a value of $v_{in} = 2.5$ mV, the following results are obtained.

$$\begin{aligned} v_{eff1} &= 0.32322 \text{ V} \\ v_{eff2} &= 0.32228 \text{ V} \\ v_{xy} &= 0.0015648 \text{ V} \end{aligned} \tag{12.79}$$

These values of effective gate voltages imply that the drain currents are

$$i_{D1} = 100.2898\ \mu\text{A} \qquad i_{D2} = 99.7104\ \mu\text{A} \tag{12.80}$$

which, in turn, imply that

$$i_{o1} = i_{D1} - 100\ \mu\text{A} = 0.2898\ \mu\text{A} \tag{12.81}$$

Such a result agrees closely with that of part 1.

However, for the case of $v_{in} = 250$ mV, we have

$$\begin{aligned} v_{eff1} &= 0.35452 \text{ V} \\ v_{eff2} &= 0.28749 \text{ V} \\ v_{xy} &= 0.18297 \text{ V} \end{aligned} \tag{12.82}$$

resulting in $i_{o1} = 20.66\ \mu$A. This result is about 30 percent lower than that expected by a linear transconductor value of 29.05 μA, as found in part 1.

3. Finally, if the gates are connected to the input signal, the preceding equations in part 2 remain the same, except for i_{xy}, which is now the sum of i_{D3} and i_{D4}, resulting in,

$$\begin{aligned} i_{xy} = {} & \mu_n C_{ox}\left(\frac{W}{L}\right)_3 \left[(v_{eff2} + v_{in})v_{xy} - \frac{v_{xy}^2}{2}\right] \\ & + \mu_n C_{ox}\left(\frac{W}{L}\right)_4 \left(v_{eff2}v_{xy} - \frac{v_{xy}^2}{2}\right) \end{aligned} \tag{12.83}$$

and so (12.77) is replaced with

$$960v_{eff2}^2 + 576[(2v_{eff2} + v_{in})v_{xy} - v_{xy}^2] = 100 \tag{12.84}$$

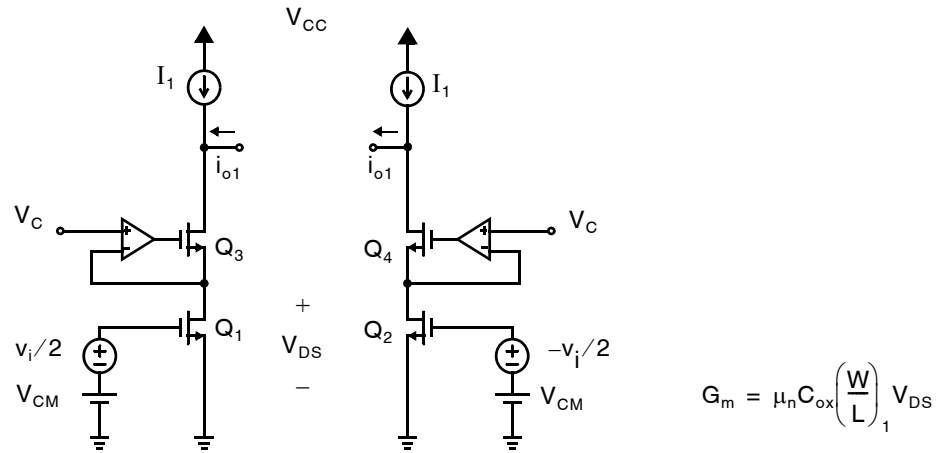


Fig. 12.20 A differential transconductor using triode transistors that have constant drain-source voltages. The control signal, V_C, adjusts the transconductance value, whereas V_{CM} is a fixed common-mode voltage. The top current pair, I_1, is set equal to the bias current through Q_1, Q_2, using a common-mode feedback circuit (not shown).

Once again, using an equation solver, we find that in the case of $v_{in} = 2.5$ mV, the results are essentially the same as in parts 1 and 2. However, in the case of $v_{in} = 250$ mV, we find that

$$\begin{aligned} v_{eff1} &= 0.36620 \text{ V} \\ v_{eff2} &= 0.27245 \text{ V} \\ v_{xy} &= 0.15625 \text{ V} \end{aligned} \tag{12.85}$$

These values lead to $i_{o1} = 28.74$ μA, which is only 1 percent below that expected by a linear transconductor value of 29.05 μA.

12.4.3 Transconductors Using Constant Drain-Source Voltages

Recall, once again, the classical model for a transistor biased in the triode region,

$$i_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{12.86}$$

We see here that if the drain-source voltage, V_{DS}, is kept constant, the drain current is linear with respect to an applied gate-source voltage (ignoring the dc offset current due to the squared V_{DS} term) [Pennock, 1985]. Unfortunately, this simple model does not account for important second-order effects (such as velocity saturation and mobility degradation) that severely limit the accuracy of the preceding equation model. However, if fully differential structures are used to cancel even-order error terms, a reasonably linear transconductor (around 50 dB) can be built using this constant-V_{DS} approach.

One possible implementation based on the linearization scheme just described is shown in Fig. 12.20. Here, Q_1 and Q_2 are placed in the triode region and their drain-source voltages are set equal to V_C through the use of Q_3, Q_4

and two extra amplifiers. The differential input signal is applied to the gates of Q_1, Q_2 and is superimposed on a constant common-mode voltage signal, V_{CM}. An alternative design here is to use an extra opamp to create a G_m-C opamp integrator [Wong, 1989].

For the circuit of Fig. 12.20, we have

$$i_{D1} = \mu_n C_{ox}\left(\frac{W}{L}\right)_1\left[\left(\frac{v_i}{2} + V_{CM} - V_{tn}\right) V_{DS} - V_{DS}^2/2\right] \tag{12.87}$$

Noting that i_{o1} is the difference between this current and I_1, which is set via a common-mode feedback signal equal to

$$I_1 = \mu_n C_{ox}\left(\frac{W}{L}\right)_1\left[(V_{CM} - V_{tn})V_{DS} - V_{DS}^2/2\right] \tag{12.88}$$

we then have

$$i_{o1} = i_{D1} - I_1 = \mu_n C_{ox}\left(\frac{W}{L}\right)_1\left(\frac{v_i}{2}\right)V_{DS} \tag{12.89}$$

Then, defining $G_m \equiv i_{o1}/v_i$ results in

$$G_m = \mu_n C_{ox}\left(\frac{W}{L}\right)_1 V_{DS} \tag{12.90}$$

Note that here the transconductance value is proportional to the drain-source voltage, and from (12.88), we see that the bias current is also approximately proportional to the drain-source voltage (if V_{DS} is small).

EXAMPLE 12.7

Consider the circuit shown in Fig. 12.20, where $\mu_n C_{ox} = 96\ \mu A/V^2$, $(W/L)_1 = (W/L)_2 = 10$, $V_{tn} = 0.8$ V, and $V_{CM} = 2$ V. Find the transconductance value when $V_C = 0.2$ V and when $V_C = 0.4$ V. For these two control-signal values, also find the bias currents through Q_1 and Q_2 (i.e., when $v_i = 0$) and the maximum input signal swing before either transistor exits the triode region (assume the classical transistor model).

Solution

Using (12.90), for $V_C = 0.2$ V, we have

$$G_m = 96\ \mu A/V^2 \times 10 \times 0.2\ V = 0.192\ mA/V$$

Also, the bias current through Q_1 or Q_2 for this V_C value can be found using (12.88):

$$I_1 = 96\ \mu A/V^2 \times 10\left[(2 - 0.8)0.2 - \frac{0.2^2}{2}\right] = 0.211\ mA$$

Finally, to remain in the triode region, the *effective* gate-source voltages of Q_1 and Q_2 must remain above $V_{DS} = 0.2$ V. In other words, their gate-source voltages must remain above $V_{DS} + V_t = 1.0$ V. Since the common-mode voltage, V_{CM}, equals 2 V, the negative portion, $v_i/2$, must not go below a level of –1 V, otherwise the transistor will enter the active region. Thus, a maximum differential input voltage, v_i, of 2 V can be applied in this case.

Similarly, for $V_C = 0.4$ V, we have twice the transconductance value of $G_m = 0.384$ mA/V. Here, the bias current is found to be $I_1 = 0.384$ mA and the maximum input signal swing is a differential voltage of $v_i = 1.6$ V. Note that doubling the drain-source voltage doubles the transconductance (as expected) but requires almost a doubling of the bias current and a loss of signal swing.

12.5 CMOS TRANSCONDUCTORS USING ACTIVE TRANSISTORS

In general, CMOS transconductors based on the use of active transistors[2] have worse linearity performance than their triode-based counterparts. However, active-based designs often have improved speed performance (assuming equal bias currents in the two cases). The linearity of active-transistor transconductors is only moderate (around 30 to 50 dB) because they rely on the square-law model of MOS transistors, which is not very accurate (particularly in short-channel processes). In addition, only the difference in output currents is ideally linear, whereas individual currents have significant second-order harmonics. Thus, second-order harmonic distortion also occurs if the difference is not taken precisely.

Key Point: *Transconductors whose gain is based upon the small-signal transconductance of active-mode transistors generally offer increased speed, but reduced linearity, compared with transconductors based upon triode devices.*

We should mention here that not all linearization techniques using active transistors are discussed. For example, one approach not discussed is a technique in which the tail current of a differential pair is adaptively biased [Nedungadi, 1984]. This technique is limited in its frequency response, and thus a designer would usually choose a triode-based design instead. In summary, in this section we present CMOS linearization techniques intended for high-speed operation and only moderate linearity.

12.5.1 CMOS Pair

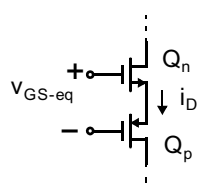
Before proceeding to look at transconductor circuits that have active transistors, it is worthwhile to introduce a two-transistor circuit known as the *CMOS pair*, as shown in Fig. 12.21 [Seevinck, 1987]. As we will show, this two-transistor pair acts as a single transistor, assuming that both transistors remain in the active region. Such an observation allows one to more easily derive some of the CMOS transconductors described later in this section.

To show that this two-transistor pair acts as a single transistor, we first assume that both transistors remain in the active region. Thus, we can write their drain currents as

$$I_D = K_n(V_{GSn} - V_{tn})^2 \qquad \text{where } V_{tn} > 0 \tag{12.91}$$

and

$$I_D = K_p(V_{SGp} + V_{tp})^2 \qquad \text{where } V_{tp} < 0 \tag{12.92}$$



$$i_D = K_{eq}\left(v_{GS\text{-}eq} - V_{t\text{-}eq}\right)^2$$

$$K_{eq} = \frac{K_n K_p}{\left(\sqrt{K_n} + \sqrt{K_p}\right)^2}$$

$$V_{t\text{-}eq} = V_{tn} - V_{tp}$$

(where $V_{tn} > 0$ and $V_{tp} < 0$)

Fig. 12.21 A CMOS pair, where both transistors are in the active region. The transistor pair is equivalent to a single transistor with an equivalent threshold voltage, $V_{t\text{-}eq}$, and device parameter, K_{eq}

2. Active transistors are those transistors that operate in the active region. The active region is also known as the pinch-off, or saturation, region. This active region for an n-channel transistor occurs when V_{GS} is greater than V_{tn} and $V_{DS} \geq V_{GS} - V_{tn} = V_{eff}$.

Rearranging these two equations in terms of their gate-source voltages gives

$$V_{GSn} = V_{tn} + \frac{1}{\sqrt{K_n}}\sqrt{I_D} \tag{12.93}$$

and

$$V_{SGp} = -V_{tp} + \frac{1}{\sqrt{K_p}}\sqrt{I_D} \tag{12.94}$$

Defining the equivalent gate-source voltage, $V_{GS\text{-}eq}$, to be the difference between the two gate voltages, we have

$$V_{GS\text{-}eq} \equiv V_{GSn} + V_{SGp} = (V_{tn} - V_{tp}) + \left(\frac{1}{\sqrt{K_n}} + \frac{1}{\sqrt{K_p}}\right)\sqrt{I_D} \tag{12.95}$$

Since this equation has the same form as those for a single transistor (equations (12.93) and (12.94)), we define the equivalent threshold voltage, $V_{t\text{-}eq}$, as

$$V_{t\text{-}eq} \equiv V_{tn} - V_{tp} \qquad \text{where } V_{tn} > 0 \text{ and } V_{tp} < 0 \tag{12.96}$$

and we define the equivalent device parameter, K_{eq}, as

$$\frac{1}{\sqrt{K_{eq}}} \equiv \frac{1}{\sqrt{K_n}} + \frac{1}{\sqrt{K_p}} \tag{12.97}$$

which can be rewritten as

$$K_{eq} = \frac{K_n K_p}{\left(\sqrt{K_n} + \sqrt{K_p}\right)^2} \tag{12.98}$$

Finally, combining (12.95), (12.96), and (12.98) and rearranging, we see that when both transistors are in their active regions, the drain current can be written as

$$i_D = K_{eq}(v_{GS\text{-}eq} - V_{t\text{-}eq})^2 \tag{12.99}$$

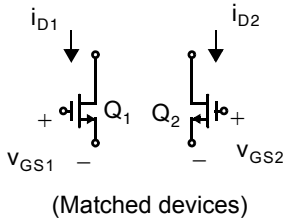
where K_{eq}, $V_{GS\text{-}eq}$, and $V_{t\text{-}eq}$ are defined as just described. Clearly, this equation is in the same form as that for a single transistor, and thus this two-transistor pair can replace single transistors in many circuits. As we will see next, this replacement can be quite useful in designing linear transconductor circuits.

12.5.2 Constant Sum of Gate-Source Voltages

One popular approach for creating an active transistor transconductor is by using a constant sum of gate-source voltages, as shown in Fig. 12.22. Assuming the two transistors are operating in the active region (i.e., the drain voltages are high and each of v_{GS1} and v_{GS2} remain greater than V_{tn}), the output differential current is given by

$$(i_{D1} - i_{D2}) = K\left(v_{GS1} + v_{GS2} - 2V_{tn}\right)(v_{GS1} - v_{GS2}) \tag{12.100}$$

Here, the *output differential current is linear if the sum of the two gate-sour ce voltages remains constant*. As we will see next, there are a variety of ways to make the sum of the gate-source voltages remain constant when applying an input signal. One important point to note here is that, although the differential current is linear, the individual drain currents are not linear. Thus, if the subtraction between currents has some error (as it certainly will), some distortion products will occur even if perfect square-law devices are obtainable.



$$K = (\mu_n/2)C_{OX}\left(\frac{W}{L}\right)$$

$$i_{D1} = K\left(v_{GS1} - V_{tn}\right)^2 \quad \text{and} \quad i_{D2} = K\left(v_{GS2} - V_{tn}\right)^2$$

$$(i_{D1} - i_{D2}) = K\left(v_{GS1} + v_{GS2} - 2V_{tn}\right)\left(v_{GS1} - v_{GS2}\right)$$

Fig. 12.22 Two-transistor circuit in which the difference in the drain currents is linear with respect to $(v_{GS1} - v_{GS2})$ if the sum of the two gate-source voltages remains constant. Transistors are matched and in the active region.

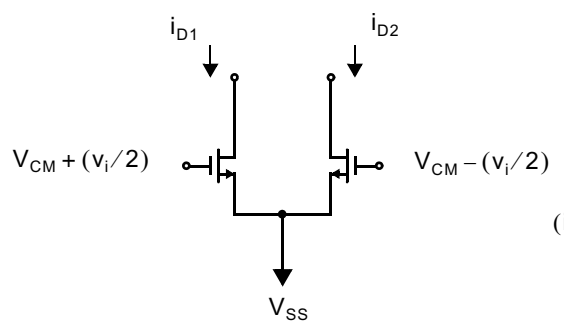
12.5.3 Source-Connected Differential Pair

One approach to maintain a constant sum of gate-source voltages is to apply a balanced differential input voltage, as shown in Fig. 12.23. Here, the differential pair is biased using the dc common-mode voltage, V_{CM}, whereas the input signal varies *symmetrically* around this common-mode voltage. Note that, from the preceding analysis, this input signal symmetry is important in maintaining linearity. However, in practice, a slightly nonbalanced input signal is not the dominant source of nonlinearity. Specifically, the linearity achievable with this approach is quite limited because MOS transistors are not well modelled by square-law behavior. In addition, it is difficult to obtain a precise difference between the two drain currents, which also causes even-order harmonics. As a result, this technique is typically limited to less than 50 dB of linearity.

To tune the transconductance of this circuit, the common-mode voltage, V_{CM}, can be adjusted. However, it should be noted that, in a short-channel process and when moderate gate-source voltages are used, carrier velocity saturation occurs and limits the transconductance variation. In addition, one should be aware that as the input common-mode voltage is adjusted, linearity performance and maximum signal swing levels change as well. One example of this approach is presented in [Snelgrove, 1992] where speeds up to 450 MHz are demonstrated in a 0.9-μm CMOS process with around a 30- to 40-dB dynamic range performance.

12.5.4 Inverter-Based

A transconductor approach similar to the source-connected differential pair is the inverter-based transconductor, shown in Fig. 12.24 [Park, 1986]. Here, two CMOS pairs, Q_{1a}, Q_{1b} and Q_{2a}, Q_{2b}, have replaced the two individual



$$(i_{D1} - i_{D2}) = 2K\left(V_{CM} - V_{SS} - V_{tn}\right)(v_i)$$

$$G_m = 2K\left(V_{CM} - V_{SS} - V_{tn}\right)$$

Fig. 12.23 A source-connected differential-pair transconductor. The differential input signal must be balanced around V_{CM} for linearization of the transistor square-law relation to occur.

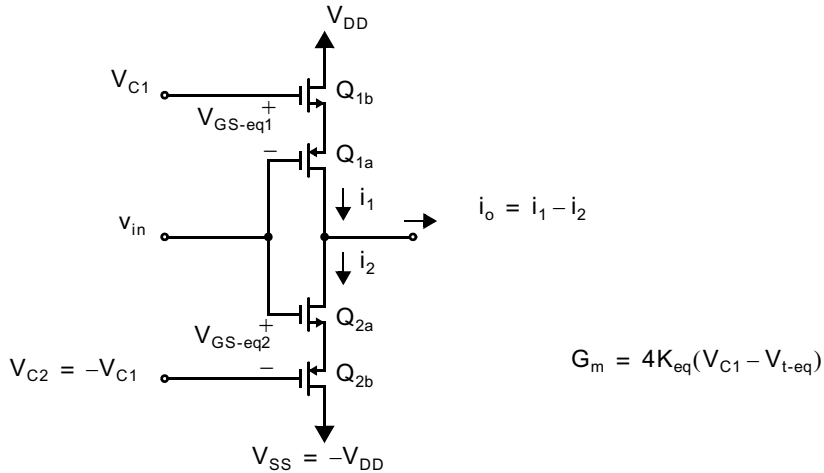


Fig. 12.24 Complementary differential-pair transconductor using CMOS pairs.

transistors in a traditional CMOS inverter, and the difference in their output currents is the overall output current, i_o. Assuming that all four transistors are in the active region and that the two CMOS pairs are matched, we can use the previous results on CMOS pairs to write

$$i_1 = K_{eq}(V_{GS\text{-}eq1} - V_{t\text{-}eq})^2 \tag{12.101}$$

and

$$i_2 = K_{eq}(V_{GS\text{-}eq2} - V_{t\text{-}eq})^2 \tag{12.102}$$

Note here that these equations are in the same form as those shown in Fig. 12.22. Thus, as in (12.100), the output current is given by

$$i_o \equiv i_1 - i_2 = K_{eq}(V_{GS\text{-}eq1} + V_{GS\text{-}eq2} - 2V_{t\text{-}eq})(V_{GS\text{-}eq1} - V_{GS\text{-}eq2}) \tag{12.103}$$

Thus, the output current is linearly related to the difference in equivalent gate-source voltages of the CMOS pairs if the sum of their equivalent gate-source voltages is a constant value. To see that a constant equivalent gate-source voltage sum occurs, we have

$$V_{GS\text{-}eq1} = V_{C1} - v_{in} \tag{12.104}$$

and

$$V_{GS\text{-}eq2} = v_{in} - V_{C2} \tag{12.105}$$

and therefore

$$V_{GS\text{-}eq1} + V_{GS\text{-}eq2} = V_{C1} - V_{C2} \tag{12.106}$$

Also, the difference in equivalent gate-source voltages is given by

$$V_{GS\text{-}eq1} - V_{GS\text{-}eq2} = V_{C1} + V_{C2} - 2v_{in} \tag{12.107}$$

Finally, in the case where we choose $V_{C2} = -V_{C1}$, we have

$$i_o = 4K_{eq}(V_{C1} - V_{t\text{-}eq})(v_{in}) \tag{12.108}$$

and the transconductance value can be varied by changing the control voltages, V_{C1} and V_{C2} (which, incidentally, are high-input impedance nodes and thus are easily adjustable). Note that, with this transconductor, matching need occur only for transistors of the same type. Specifically, the n-channel transistors should match and the p-channel transistors should match, but there is no need to match n- and p-channel transistors to each other.

12.5.5 Differential-Pair with Floating Voltage Sources

Another approach to maintain a constant sum of gate-source voltages is by using two floating dc voltage sources, as shown in Fig. 12.25 [Nedungadi, 1984]. Writing a voltage equation around the loop, we have

$$v_{GS1} - (V_x + V_{tn}) + v_{GS2} - (V_x + V_{tn}) = 0 \tag{12.109}$$

and thus,

$$v_{GS1} + v_{GS2} = 2(V_x + V_{tn}) \tag{12.110}$$

As a result, this circuit maintains a constant sum of gate-source voltages even if the applied differential signal is not balanced. Also, writing two equations that describe the relationship between v_1 and v_2, we have

$$v_1 - v_{GS1} + V_x + V_{tn} = v_2 \tag{12.111}$$

and

$$v_2 - v_{GS2} + V_x + V_{tn} = v_1 \tag{12.112}$$

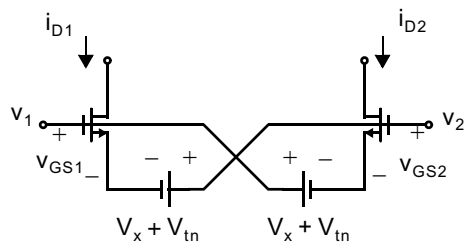
Subtracting (12.111) from (12.112) and rearranging, we obtain

$$v_{GS1} - v_{GS2} = 2(v_1 - v_2) \tag{12.113}$$

Thus, the difference between the input voltages is equal to one half the difference between the gate-source voltages. Finally, the differential output current is found by combining (12.100), (12.110), and (12.113) to give

$$(i_{D1} - i_{D2}) = 4KV_x(v_1 - v_2) \tag{12.114}$$

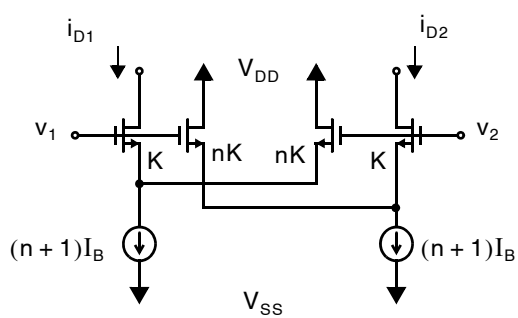
One simple way to realize the floating voltage sources of Fig. 12.25 is to use two source followers, as shown in Fig. 12.26 [Nedungadi, 1984]. The transistors labelled nK are n times larger than the other two transistors, and thus



$$(i_{D1} - i_{D2}) = 4KV_x(v_1 - v_2)$$

$$G_m = 4KV_x$$

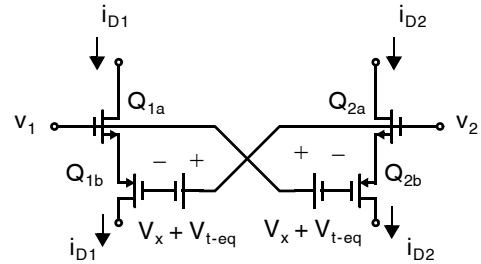
Fig. 12.25 A linearized differential pair using floating voltage sources.



$$(i_{D1} - i_{D2}) = \left(\frac{n}{n+1}\right) 4\sqrt{KI_B}(v_1 - v_2)$$

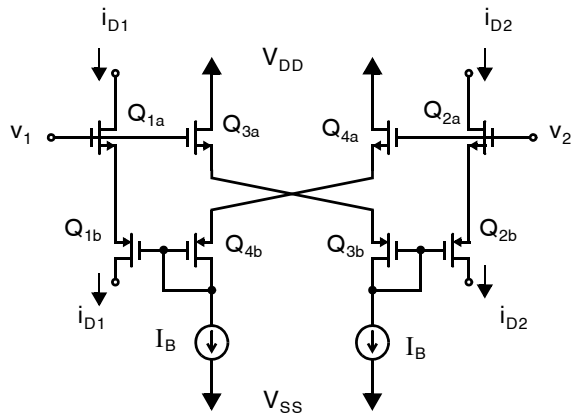
$$G_m = \left(\frac{n}{n+1}\right) 4\sqrt{KI_B}$$

Fig. 12.26 A linear transconductor for large n.



$$(i_{D1} - i_{D2}) = 4K_{eq}V_x(v_1 - v_2)$$

Fig. 12.27 A linearized differential pair in which the floating voltage sources need drive only the transistor gates.



$$(i_{D1} - i_{D2}) = 4\sqrt{K_{eq}I_B}(v_1 - v_2)$$
$$G_m = 4\sqrt{K_{eq}I_B}$$

Fig. 12.28 A transconductor that uses CMOS pairs and floating voltage sources.

they act as source followers when n is large (typically, when n is 5 or greater). If n is not large, the bias current through the two source followers changes as the input signal is applied and linearity is worsened. A disadvantage of this circuit is that a large dc bias current passes through the two source followers. However, if only moderate linearity is desired and the input signal is relatively small (say, 500 mV), then this approach can be quite useful.

Since it is difficult to realize a floating voltage source that drives the source of another transistor, another approach is to modify the circuit of Fig. 12.25 such that the floating voltage sources need drive only the transistor gates. Such a modification is possible by replacing the single transistors shown in Fig. 12.25 with their CMOS pairs, as shown in Figure 12.27. Note that the voltage sources must now be larger than $V_{t\text{-}eq}$ and must include a controllable V_x voltage. Fortunately, such circuits are easily accomplished by using two more CMOS pairs, as shown in Fig. 12.28, where the pairs Q_{3a}, Q_{3b} and Q_{4a}, Q_{4b} have been added together with two adjustable current sources [Seevinck, 1987]. Referring back to (12.99), we see that V_x is given by

$$V_x = \frac{1}{\sqrt{K_{eq}}}\sqrt{I_B} \tag{12.115}$$

where, for simplicity, we have assumed that all CMOS pairs are matched. As a result, the differential output current is given by

$$(i_{D1} - i_{D2}) = 4\sqrt{K_{eq}I_B}(v_1 - v_2) \tag{12.116}$$

Thus, we see that the output current can be tuned by adjusting the dc bias current, I_B. Note that the transconductance coefficient is proportional to the square root of I_B and thus a fourfold increase in current results in only a factor of 2 coefficient increase.

Before leaving this circuit, a couple of points are worth mentioning. First, note that the output currents from this circuit are available at both the upper pair, Q_{1a}, Q_{2a}, as well as at the lower pair, Q_{1b}, Q_{2b}. Such flexibility may be useful in the remaining transconductor circuitry. Second, note that V_{SS} must be at least $V_{tn} + |V_{tp}|$ below the most negative input signal swing of either v_1 or v_2. If we assume v_1 and v_2 have ±1-V signals associated with them around ground, then, for $V_{tn} = -V_{tp} = 1$V, V_{SS} would have to be less than –3 V. Thus, assuming we have symmetrical power supplies, a 6-V power supply would be needed for only a 2-V signal swing.

12.5.6 Bias-Offset Cross-Coupled Differential Pairs

Another approach to realize a transconductor with active transistors is to use two cross-coupled differential pairs where the input into one pair is intentionally voltage offset [Wang, 1990]. One implementation of this approach is shown in Fig. 12.29, where the differential pair Q_1, Q_2 is cross coupled with the bias-offset differential pair Q_3, Q_4. Here, we assume all transistors are in the active region and are matched (i.e., we assume all transistors have equal device parameters of K and equal threshold voltages of V_{tn}).

To find the differential output current, we first find

$$i_1 = K(v_1 - V_x - V_{tn})^2 + K(v_2 - V_B - V_x - V_{tn})^2 \tag{12.117}$$

and

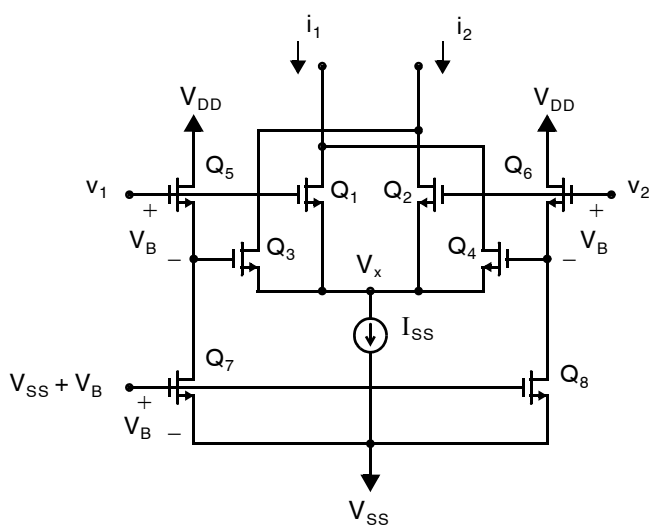
$$i_2 = K(v_2 - V_x - V_{tn})^2 + K(v_1 - V_B - V_x - V_{tn})^2 \tag{12.118}$$

Now finding the difference, $i_1 - i_2$, we have

$$(i_1 - i_2) = 2KV_B(v_1 - v_2) \tag{12.119}$$

Thus, the output differential current is linear with respect to the differential input voltage (as expected) and the transconductance value is proportional to V_B. Note, however, that the bias currents through Q_5, Q_6, Q_7, and Q_8 are all square-law related to V_B. Thus, the transconductance value is proportional to the square root of the changing bias current, as in the other active transistor CMOS transconductors. Also, note that the bias current, I_{SS}, does not affect the transconductance value but does determine the maximum (or minimum) output current available.

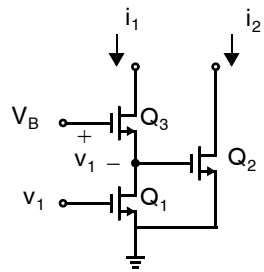
Finally, note that this offset-bias linearization technique can also be realized in a single-ended version with only a few transistors, as shown in Fig. 12.30 [Bult, 1987]. The analysis is left to the reader.



$$(i_1 - i_2) = 2KV_B(v_1 - v_2)$$

$$G_m = 2KV_B$$

Fig. 12.29 A bias-offset cross-coupled transconductor.



$$i_1 - i_2 = K\left(V_B - 2V_{tn}\right)\left(2v_1 - V_B\right)$$

Fig. 12.30 A single-ended offset-bias transconductor.

12.6 BIPOLAR TRANSCONDUCTORS

There are two main approaches for realizing bipolar transconductors. One approach is to use a fixed transconductor cascaded with a gain cell. The fixed transconductor is typically a differential pair linearized through resistor degeneration. The gain cell allows scaling of the output current by varying the ratio of two current sources. The other approach is to use a differential input stage with multiple inputs, where the transistors are scaled such that improved linearity occurs. In this case, the transconductance value can be tuned by varying the bias current through the input stage.

12.6.1 Gain-Cell Transconductors

Key Point: *Bipolar transconductors based on fixed resistors may be made tunable by cascading them with an adjustable-gain cell.*

The bipolar transconductors in Fig. 12.12 and Fig. 12.14 have a fixed transconductance value equal to approximately $1/R_S$. However, to accommodate tuning, variable transconductance values are required. One common approach for creating tunable transconductors that have a fixed transconductance input stage is to use a gain cell as described in Section 11.5. Recall that in a gain cell, the output current is equal to a scaled version of the input current, where the scaling factor is determined by the ratio of two dc bias currents. In addition, the circuit is highly linear.

A tunable transconductor that uses a gain cell is shown in Fig. 12.31. Here, the voltage sources, V_{LS}, represent voltage level shifters and could be realized using the circuit shown in Fig. 12.32. By combining the relationships given for the fixed transconductor and the gain cell, specifically, (12.40) and (11.17), then the output current is given by

$$i_o = \frac{v_i I_2}{R_E I_1} \tag{12.120}$$

resulting in a transconductance value equal to

$$G_m = \left(\frac{1}{R_E}\right)\left(\frac{I_2}{I_1}\right) \tag{12.121}$$

Thus, the transconductance value is proportional to the ratio of I_2/I_1.

It is also possible to place the gain cell below the input differential pair, as shown in Fig. 12.33 [Wyszynski, 1993; Moree, 1993]. Note that while the previous designs have $2r_e$ in series with the degeneration resistor, R_E, this circuit has $4r_e$ in series with R_E (that of $Q_1 - Q_4$), which might affect the distortion performance. However, the gain cell shown has less distortion due to finite β effects and has therefore been referred to as a *beta-immune type-A cell* [Gilbert, 1990].

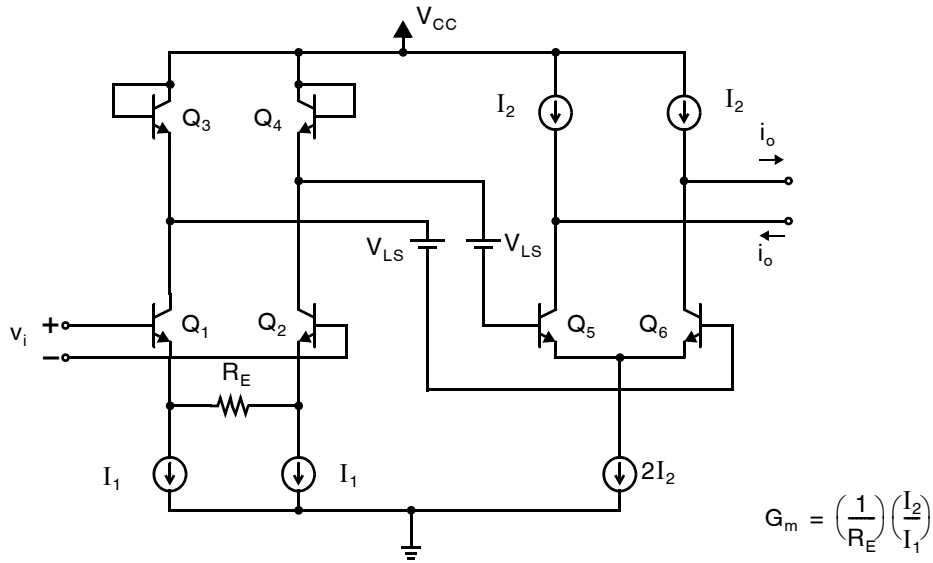


Fig. 12.31 A typical gain-cell transconductor.

12.6.2 Transconductors Using Multiple Differential Pairs

Quite a different approach for realizing a bipolar transconductor is by partially linearizing the input stage through the use of multiple differential pairs. To see how this linearization is achieved, consider a simple differential pair, as shown in Fig. 12.34(*a*). Ignoring base currents, we saw in Section 8.5.3 that the collector current i_{C2} is given by

$$i_{C2} = \frac{I_1}{1 + e^{v_i/V_T}} \tag{12.122}$$

This current is plotted in Fig. 12.34(*b*), where we note that the curve is relatively linear near $v_i = 0$. It is well known that the slope at this point equals the transconductance of this differential Ginput stage and is given by

$$G_m = \frac{1}{2r_e} = \frac{I_1}{4V_T} \tag{12.123}$$

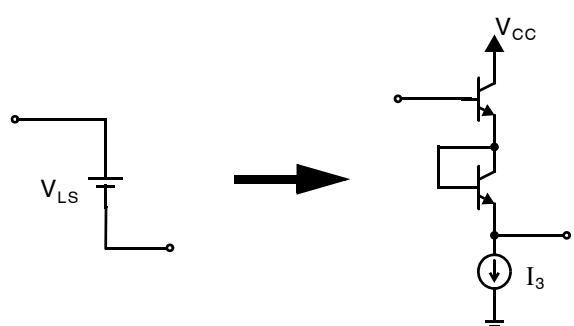


Fig. 12.32 A possible realization of the level shifters.

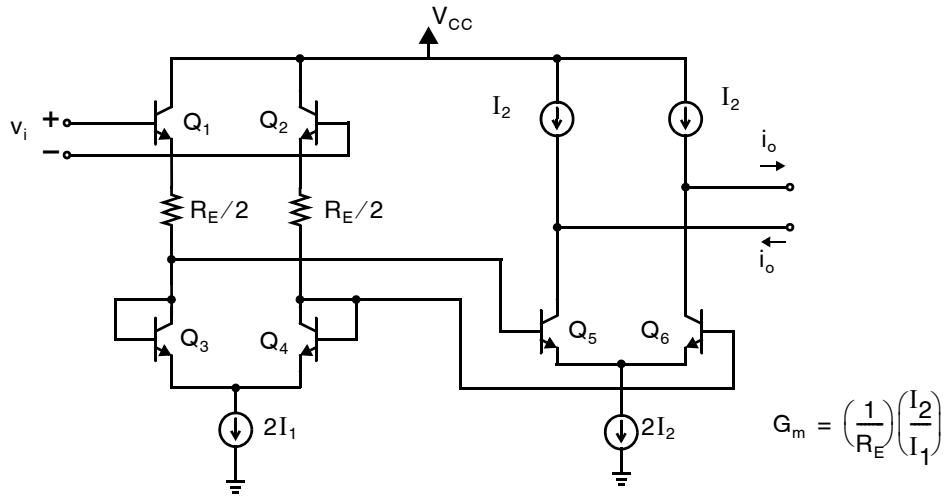


Fig. 12.33 A gain-cell transconductor with the gain cell below the input differential pair.

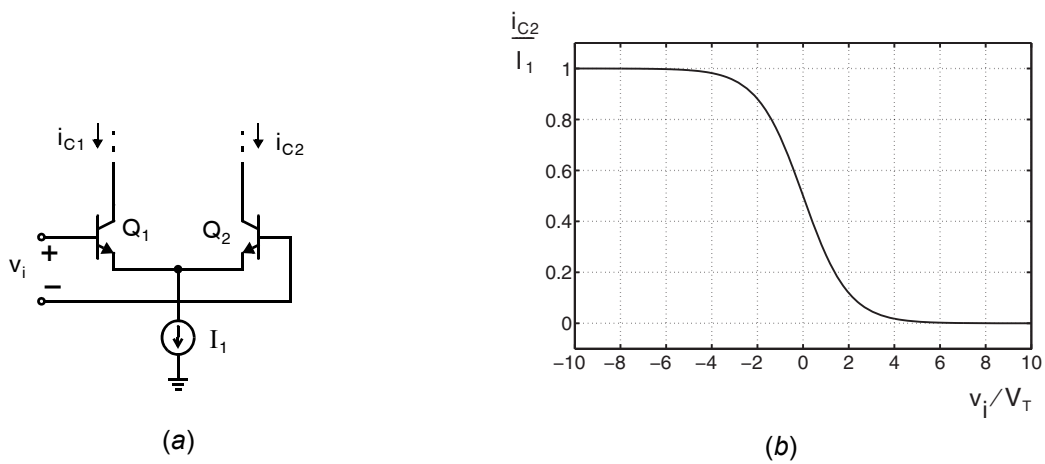


Fig. 12.34 (*a*) Differential pair. (*b*) Current output versus input voltage.

Thus, we see that the transconductance is proportional to the bias current, I_1. Unfortunately, this two-transistor input stage has a very limited input range when used as a linear transconductor. In fact, when v_i becomes greater than around 32 mV_{pp} (i.e., $v_i > \pm 16$ mV), the total harmonic distortion of the output current becomes greater than 1 percent (see Example 12.8). We now discuss a method to use parallel differential pairs to increase the linear input range [Calder, 1984].

Consider the input stage shown in Fig. 12.35(*a*), where two differential pairs are connected in parallel but with a dc offset voltage applied to each. The currents of the right side transistors in each transistor pair are shown in Fig. 12.35(*b*). If the dc offset voltages, v_1, are chosen carefully, then the two current curves will partially linearize each other when they are added together, as shown in Fig. 12.35(*c*). In this case, v_1 is chosen to be equal to $1.317V_T$ to maximize the linear input range [Tanimoto, 1991]. As a result of this linearization technique, the input range can now be approximately three times that of the original differential pair or, equivalently, 96 mV_{pp}, and achieve the same distortion performance.

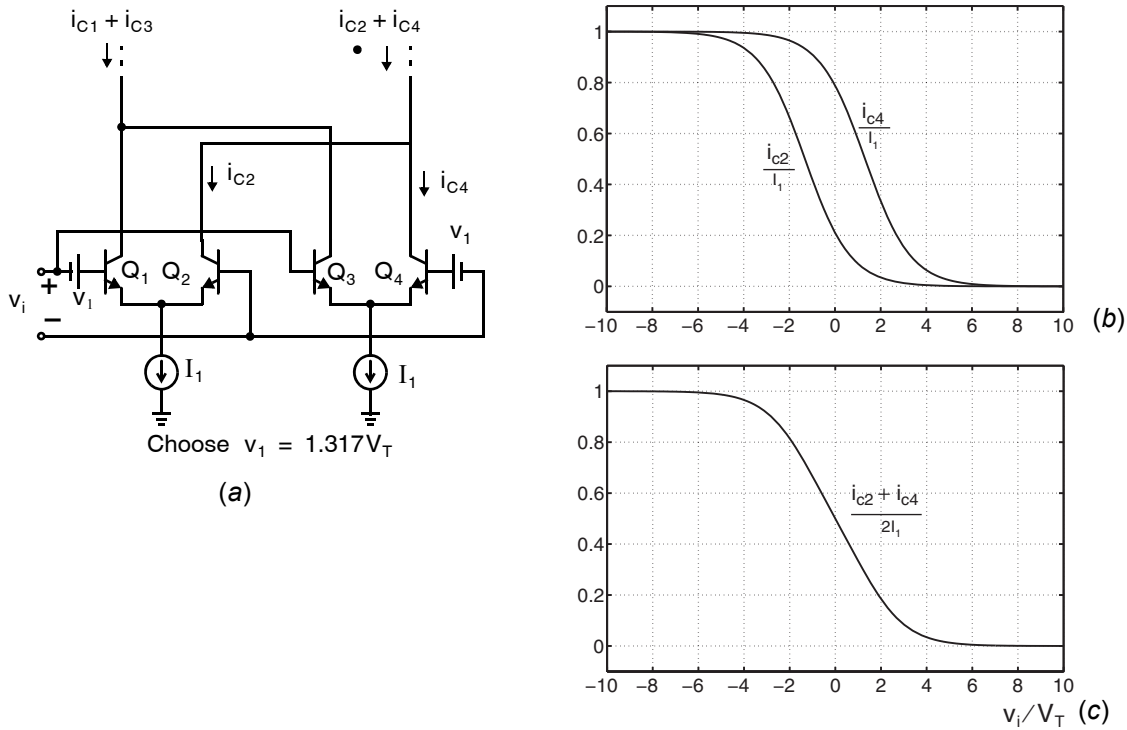


Fig. 12.35 (*a*) A linearized differential input; (*b*) individual right-side currents; (*c*) total linearized

To reduce circuitry, the two dc offset voltage sources can be eliminated by sizing the differential pair of transistors appropriately. Specifically, assuming that $v_i = 0$, then when $v_1 = 1.317V_T$, the collector currents of Q_2 and Q_4 are given by

$$i_{C2} = \frac{I_1}{1 + e^{1.317}} = 0.2113I_1 \tag{12.124}$$

and

$$i_{C4} = i_{C1} = \frac{I_1}{1 + e^{-1.317}} = 0.7887I_1 \tag{12.125}$$

If a two-transistor differential pair has unequal-sized transistors such that the same two currents occur for $v_i = 0$, then this unequal-sized differential pair will behave the same as that of an equal-sized differential pair with a dc offset applied. To determine the sizing of the unequal-sized differential pair, define k to be the ratio of the base-emitter area of Q_1 with respect to Q_2. Thus, for this unequal-sized case, when $v_i = 0$, we have

$$i_{C1} = kI_s e^{(v_{be}/V_T)} = 0.7887I_1 \tag{12.126}$$

and

$$i_{C2} = I_s e^{(v_{be}/V_T)} = 0.2113I_1 \tag{12.127}$$

Dividing (12.126) by (12.127), we obtain

$$k = 3.73 \tag{12.128}$$

Thus, we should make Q_1 3.73 times larger than Q_2, and similarily Q_4 should be 3.73 times larger than Q_3. Such transistor sizing achieves the same result as the dc offset voltages. This result should make intuitive sense

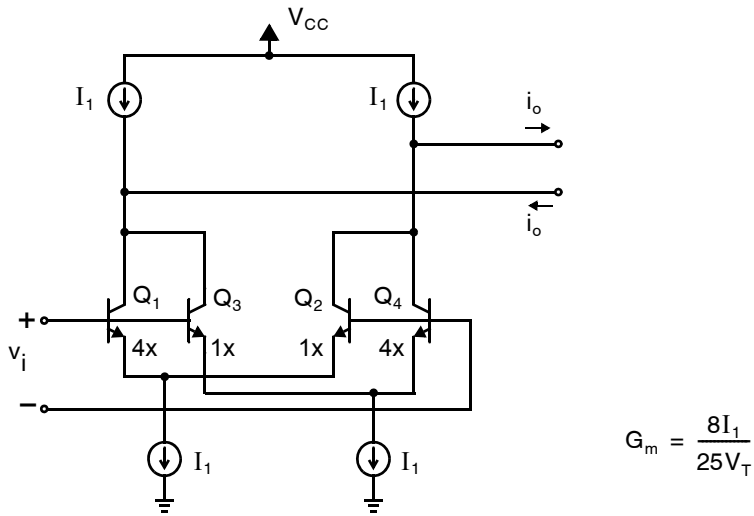


Fig. 12.36 Using transistor sizing to create a linearized differential transconductor.

since we require that the ratio of the two currents be $0.7887/0.2113$ when their base-emitter voltages are equal; this requirement results in the area ratio. Typically, this area ratio is set to 4 for practical reasons, and the final linearized transconductor is realized as shown in Fig. 12.36. Thus, for this circuit with $v_i = 0$, the currents through Q_1 and Q_2 are $0.8I_1$ and $0.2I_1$, respectively.

To determine the transconductance of the circuit of Fig. 12.36, we note that G_m equals the sum of the transconductance of the two differential pairs, resulting in

$$G_m = \frac{1}{r_{e1} + r_{e2}} + \frac{1}{r_{e3} + r_{e4}} \tag{12.129}$$

Also, at $v_i = 0$, we have

$$r_{e1} = r_{e4} = \frac{V_T}{0.8I_1} \tag{12.130}$$

and

$$r_{e2} = r_{e3} = \frac{V_T}{0.2I_1} \tag{12.131}$$

Combining (12.129) to (12.131), we have

$$G_m = \frac{8I_1}{25V_T} \tag{12.132}$$

which is 28 percent greater than the transconductance value for the transistor pair shown in Fig. 12.34(*a*). However, this comparison is not fair since two differential pairs and twice the current are used in this linearized approach. If two differential pairs of Fig. 12.34(*a*) are used together, then their total transconductance would be $G_m = I_1/(2V_T)$, implying that this linearized approach has a 36 percent smaller transconductance. Also note that, as in the two-transistor pair case, this transconductance value is also proportional to the bias current, I_1, and thus no gain cell is required.

This approach was ued in a 2- to 10-MHz programmable filter intended for disk-drive applications in [De Veirman, 1992]. This same technique can be extended to a larger number of input stages to further improve the

linear input range [Tanimoto, 1991]. Using four differential input pairs in parallel, the linear input range was reported to be 16 times better than that for a single differential pair.

EXAMPLE 12.8

Consider the single differential pair shown in Fig. 12.34(*a*), where $I_1 = 2$ mA. Find the percentage error between an ideal linear transconductor and the true output current when $v_i = 16$ mV.

Solution

Using (12.123), we have

$$G_m = \frac{I_1}{4V_T} = 19.2 \text{ mA/V} \tag{12.133}$$

Thus, for an ideal linear transconductor, the output current for a 16-mV input voltage would be

$$i_o = G_m \times 16 \text{ mV} = 0.307 \text{ mA} \tag{12.134}$$

The true output current can be found from (12.122) as

$$I_{C2} = \frac{2 \text{ mA}}{1 + e^{16/26}} = 0.702 \text{ mA} \tag{12.135}$$

$$I_{C1} = \frac{2 \text{ mA}}{1 + e^{-16/26}} = 1.298 \text{ mA} \tag{12.136}$$

implying that the change in current from the bias current of 1 mA is 0.298 mA. Thus, the percentage error between the true output current and the ideal linear transconductor output is

$$\frac{0.298 - 0.307}{0.307} \times 100 = -3\% \tag{12.137}$$

In this case, the deviation is three-percent worst case, which results in about a one-percent total-harmonic-distortion (THD) measurement (see Section 9.5.1).

EXAMPLE 12.9

Consider the linearized differential pairs shown in Fig. 12.35(*a*), where $I_1 = 2$ mA. Find the percentage error between an ideal linear transconductor and the true output current when $v_i = 16$ mV.

Solution

First, we need to derive a formula for G_m in a manner similar to that for (12.132), where

$$r_{e1} = r_{e4} = \frac{V_T}{0.7887 I_1} \tag{12.138}$$

$$r_{e2} = r_{e3} = \frac{V_T}{0.2113 I_1} \tag{12.139}$$

Thus, the transconductance is given by

$$G_m = \frac{1}{r_{e1} + r_{e2}} + \frac{1}{r_{e3} + r_{e4}} = \frac{0.333 I_1}{V_T} = 25.6 \text{ mA/V} \tag{12.140}$$

For an ideal linear transconductor, the output current for a 16-mV input voltage would result in

$$i_o = G_m \times 16 \text{ mV} = 0.4102 \text{ mA} \tag{12.141}$$

To calculate the true output current, we first note that

$$v_1 = 1.317 V_T = 34.24 \text{ mV} \tag{12.142}$$

and since $v_i = 16$ mV, we use $v_i + v_1$ to find I_{C1}, resulting in

$$I_{C1} = \frac{2 \text{ mA}}{1 + e^{-50.24/26}} = 1.747 \text{ mA} \tag{12.143}$$

$$I_{C2} = 2 \text{ mA} - I_{C1} = 0.253 \text{ mA} \tag{12.144}$$

For I_{C4}, we use $v_1 - v_i$ to obtain

$$I_{C4} = \frac{2 \text{ mA}}{1 + e^{-18.24/26}} = 1.337 \text{ mA} \tag{12.145}$$

$$I_{C3} = 2 \text{ mA} - I_{C4} = 0.663 \text{ mA} \tag{12.146}$$

The true output is then equal to

$$I_{C1} + I_{C3} = 2.41 \text{ mA} \tag{12.147}$$

$$I_{C2} + I_{C4} = 1.59 \text{ mA} \tag{12.148}$$

implying that the change in current from the bias current of 2 mA is 0.4099 mA. Thus, the percentage error between the true output current and the ideal linear transconductor output is

$$\frac{0.4099 - 0.4102}{0.4102} \times 100 = -0.1\% \tag{12.149}$$

Note that this error is over ten times better than with a simple differential pair.

12.7 BICMOS TRANSCONDUCTORS

In this section, we discuss some of the reported approaches for realizing BiCMOS transconductors and filters.

12.7.1 Tunable MOS in Triode

One approach for realizing a BiCMOS integrator is shown in Fig. 12.37 [Laber, 1993]. Here, linearization and tuning are accomplished by using transistors Q_1 through Q_4, which are all assumed to be in the triode region. As we saw in Section 12.4, the currents through these transistors are ideally linear if their drain-source voltages are kept constant. Although in practice these currents have rather large second-order distortion, fortunately even-order distortion products are significantly reduced in a fully differential structure. The drain-source voltages of Q_1 through Q_4 are controlled through the use of the two emitter followers, Q_5, Q_6. This simple voltage control is one of the benefits of using a BiCMOS process since a MOS source follower would have too much voltage variation. Note that additional inputs into this integrator are added as simply two extra transistors (a two-input integrator is shown). Various coefficient ratios can be obtained by scaling the sizes of the input transistors. However, because of the shown tuning circuit connection, all coefficients in one integrator will scale together as the control voltage is varied. Note also that a G_m-C opamp integrator is used to reduce the effects of parasitic capacitance.

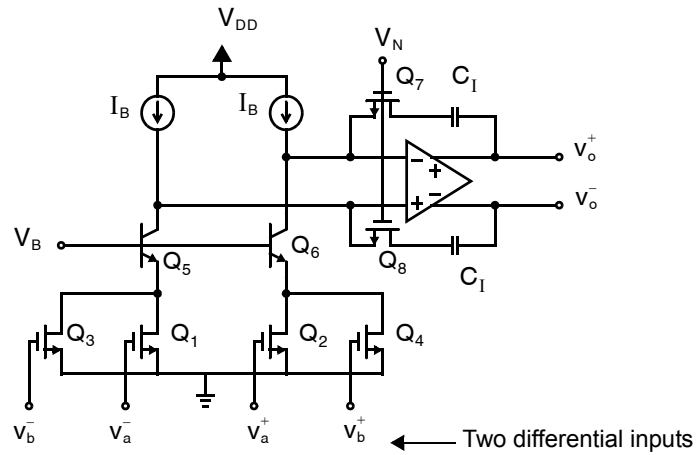


Fig. 12.37 A fully differential two-input BiCMOS integrator using MOS transistors in triode, Q_1 through Q_4 G_m-C opamp integration is used to reduce parasitic capacitance effects. Note that, in addition to a common-mode feedback circuit for the output (not shown), the common-mode voltage at the input nodes of the opamp must also be set.

Also, transistors Q_7, Q_8 are included to realize lead compensation. Finally, it should be mentioned that a common-mode feedback circuit is required at the output nodes of the differential opamp. Also, some method is needed for setting the common-mode voltage at the opamp inputs. The BiCMOS filter in [Laber, 1993] is built in a 1.5-μm, 4-GHz BiCMOS technology and operated around 20 MHz using a 5-V supply. The transconductor demonstrated moderate linearity of around 50 dB when processing 2-V_{pp} differential output signals, but no mention is made of any intermodulation results or the frequency of the test signals.

12.7.2 Fixed-Resistor Transconductor with a Translinear Multiplier

Another approach that has been successfully used to realize BiCMOS filters that have very high linearity has been to use highly linear transconductors realized using fixed resistors and then achieve tuning by following the transconductor with a translinear multiplier. As we saw in Chapter 11, the output of a translinear multiplier is linear with respect to one of its input currents, and its gain coefficient can be scaled as the ratio of two other currents in either a positive or negative fashion.

A translinear multiplier can be used for realizing a low-distortion BiCMOS integrator, as shown in Fig. 12.38 [Willingham, 1993]. Here, the output currents from fixed-value transconductors are applied to a translinear multiplier, such that tuning can be accomplished while maintaining high linearity. The outputs of the translinear multiplier are applied to a opamp-based integrator to maintain high linearity. In this implementation, the input impedance of the fully differential opamp is kept low such that a fixed voltage occurs at the input nodes of the opamp, thereby setting the input common-mode voltage. Phase-lead resistors are also used in series with the integration capacitors (not shown) to improve the integrator's phase response.

The fixed transconductors are realized using resistor degeneration and local feedback, as shown in Fig. 12.39. Fig. 12.39(*a*) shows that local feedback around the two shown opamps force their input nodes to be equal, and therefore the applied input voltage appears across R_x. A simplified circuit implementation of this approach is shown in Fig. 12.39(*b*), which shows that the opamps are implemented with only a PMOS transistor and current source.

This integrator was used to build a seventh-order low-pass filter in a 2-μm, 2.5-GHz BiCMOS process that featured thin-film resistors. The filter passband edge was tunable from 7.5 MHz to 10 MHz. The results of this low-distortion BiCMOS filter are quite impressive; third-order intermodulation products remained below 65 dB,

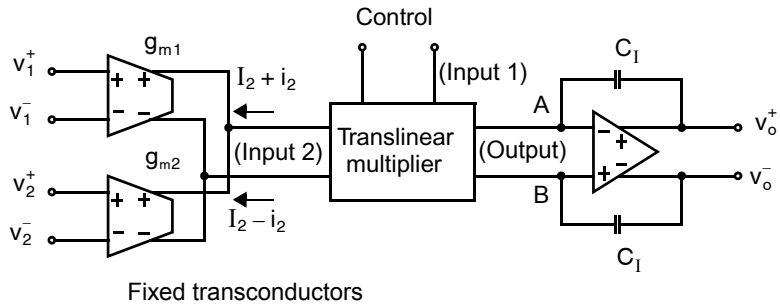


Fig. 12.38 A two-input BiCMOS integrator using fixed resistor-degenerated transconductors together with a translinear multiplier for tuning. The common-mode voltage at nodes A and B was set by using a low-input-impedance opamp.

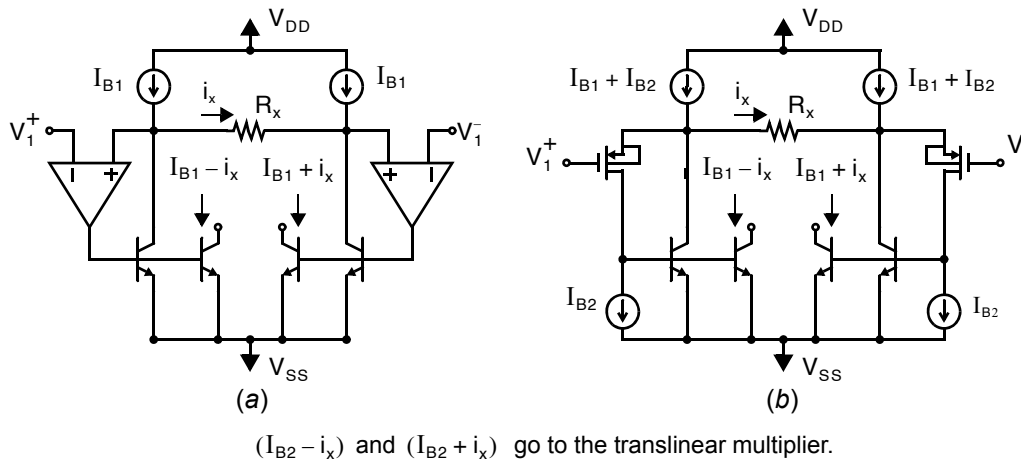


Fig. 12.39 A fixed BiCMOS transconductor: (*a*) the two local feedbacks keep the input voltage applied across R_x; (*b*) circuit implementation.

with respect to the input signal, when measured using two-tone inputs near 4 MHz. Finally, it should be mentioned that the circuit also uses a center tap on each of the R_x resistors to set the common-mode voltage of the preceding output. With this simple addition, much of the traditional common-mode circuitry is eliminated since it is necessary only to compare the center tap voltage with a dc reference signal and then feed back the appropriate control signal.

Similar approaches, in which a highly linear transconductor that has local feedback is followed by a current attenuator, are also possible using ordinary CMOS and are now gaining in popularity [Chang, 1995].

12.7.3 Fixed Active MOS Transconductor with a Translinear Multiplier

A BiCMOS integrator intended for applications that require high-speed operation, wide tuning range, and only moderate linearity is shown in Fig. 12.40 [Shoval, 1993]. In this circuit, the common-mode voltage, V_{CM}, is set to a constant value (it is set by a common-mode feedback circuit); the result is that the differential currents of Q_1, Q_4 and Q_2, Q_3 are linearly related to the balanced differential voltage (as discussed in Section 12.5). This linear voltage-to-current conversion remains at a fixed value (resulting in a relatively constant linearity performance and

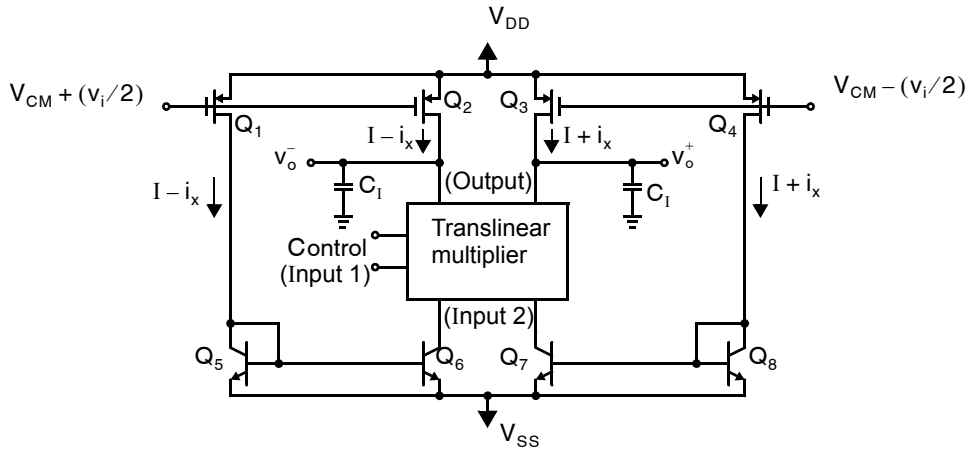


Fig. 12.40 A BiCMOS integrator using MOS transistors in the active region for linearization, and a translinear multiplier for tuning. The common-mode feedback circuit is not shown.

input swing level), and transconductance tuning is performed with the aid of a translinear multiplier. Note, however, that two pairs of MOS transistors are performing voltage-to-current conversion on the balanced input signal. The reason for using two pairs is to maximize the largest possible transconductance of the overall circuit for a given bias current. Specifically, if the multiplier is tuned to send all of i_{C6}, i_{C7} through to v_o^+, v_o^-, respectively, then the signal currents add together and the overall transconductance is equal to twice that which occurs for a single set of MOS transistors (without an increase in bias current). Alternatively, if the multiplier is tuned to send all of i_{C6}, i_{C7} through to v_o^-, v_o^+, respectively, then the currents subtract, causing a zero overall transconductance. Finally, if the translinear multiplier is tuned to cause a zero output (i.e., to cause two constant dc currents at the multiplier outputs), the overall transconductance is equal to that caused by transistors Q_2, Q_3, which is half the maximum value. Thus, the use of two pairs on the input signal doubles the maximum possible transconductance at the cost of being able to tune the transconductance only from zero to the maximum value. In other words, the overall transconductance value cannot change signs by changing the multiplier control signal. Of course, both positive and negative transconductance values can be obtained by cross-coupling the output signals.

This BiCMOS integrator is used to realize a second-order pulse-shaping filter operating on 100-Mb/s data. Although the filter has a wide tuning range of approximately 10:1, it achieves only a moderate spurious-free dynamic range level of 35 dB.

12.8 ACTIVE RC AND MOSFET-C FILTERS

Another technique for the realization of analog integrated filters is active RC or MOSFET-C filters. In both techniques, integration of current is performed upon capacitors connected in feedback around a high-gain amplifier, as opposed to transconductance-C filters which use grounded capacitors to integrate current. This is sometimes called Miller integration, since like the Miller compensation capacitor in a two-stage opamp, integrating capacitors are connected around high-gain amplifiers. Miller integration improves linearity, but the need for a high gain-bandwidth product in the amplifier makes active RC and MOSFET-C filters generally slower than their G_m-C counterparts. Also, they are further reduced in speed in a CMOS technology because opamps capable of driving resistive loads are required as opposed to only capacitive loads in G_m-C filters. However, they are particularly attractive in BiCMOS technologies where high-transconductance opamps are available.

12.8.1 Active RC Filters

Active RC filters are well established in both discrete component and integrated analog filter design. Recall that the basic building blocks of analog filters are the integrator, summer, and gain stage. A basic active RC integrator is shown in single-ended form in Fig. 12.41(*a*). Assuming an ideal opamp, a virtual ground is maintained at V_x. Hence, voltage-to-current conversion is performed by the resistor R. The gain of this conversion is determined by the resistor value, which unfortunately will vary under process and temperature variations. However, the use of a passive resistor provides excellent linearity. The resulting current is the integrated upon the capacitor C_I and the inverted output appears at v_o.

Key Point: *Active RC filters employ resistors connected to virtual ground nodes to perform voltage-to-current conversion, and then integrate the resulting currents upon Miller capacitors, connected in feedback around high-gain amplifiers.*

In Fig. 12.41(*b*), summing is incorporated into the integrator. Its fully-differential implementation is shown. The resistors $R_{p1,\,p2,\,n1,\,n2}$ convert the differential voltage inputs into differential currents. The currents $i_{p1,\,p2}$ and $i_{n1,\,n2}$ are then summed by shunting them at the virtual ground nodes V_x. The resulting total current is integrated upon the capacitors C_I. If pure gain and/or summartion is required without integration, the capacitors may be replaced by resistors in feedback around the opamp, resulting in the familiar inverting opamp configuration.

EXAMPLE 12.10

Consider the basic active RC integrator of Fig. 12.41(*a*), but assume the opamp has a first-order open-loop response characterized by a finite dc gain of A_0 and unity-gain frequency ω_{ta}. What is the resulting frequency response of the integrator?

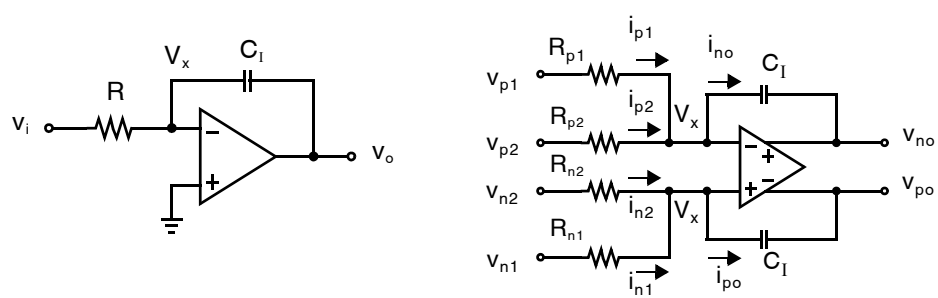
Solution

The frequency response is the expected ideal value of $-1/(sRC)$ multiplied by $L(s)/(1 + L(s))$ where $L(s)$ is the open loop response, being the product of the opamp's open-loop frequency response,

$$A(s) = \frac{A_0}{1 + \dfrac{A_0 s}{\omega_{ta}}} \tag{12.150}$$

and the feedback network's frequency response

$$\beta(s) = \frac{sRC}{1 + sRC} \tag{12.151}$$



(*a*) (*b*)

Fig. 12.41 (*a*) A single-ended active RC integrator. (*b*) A two-input active-RC fully differential integrator.

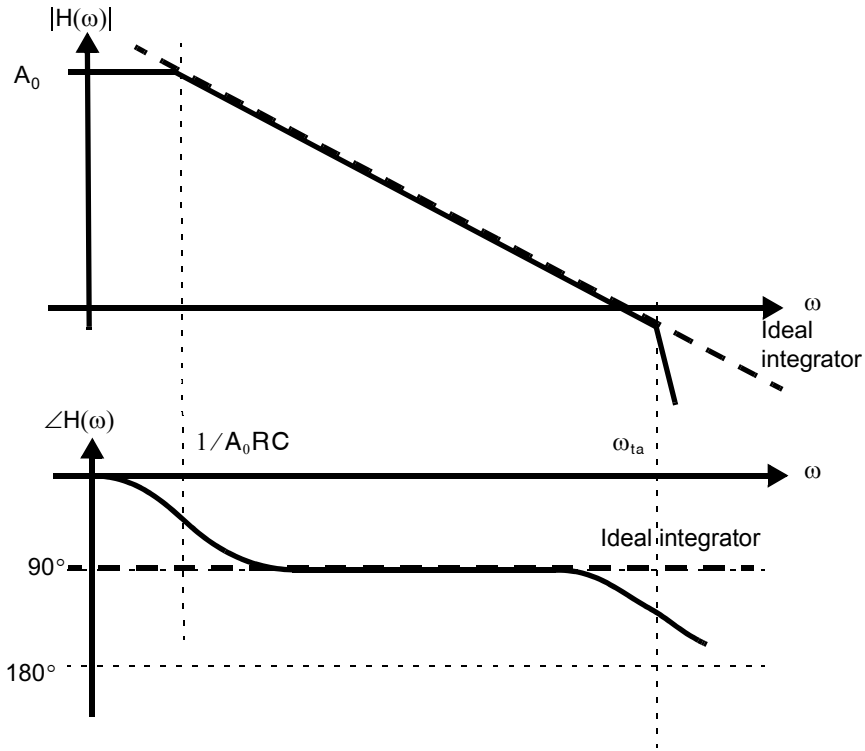


Fig. 12.42 The Bode plot of an active RC integrator with finite opamp bandwidth and unity-gain frequency.

Using (12.150) and (12.151) allows us to find the nonideal integrator's actual frequency response.

$$H(s) = -\frac{1}{(sRC)} \cdot \frac{\left(\frac{A_0}{1 + A_0 s/\omega_{ta}}\right)\left(\frac{sRC}{1 + sRC}\right)}{1 + \left(\frac{A_0}{1 + A_0 s/\omega_{ta}}\right)\left(\frac{sRC}{1 + sRC}\right)} \tag{12.152}$$

Making the assumption $\omega_{ta} >> 1/RC$, (12.152) may be manipulated to show

$$H(s) \cong -\frac{A_0}{(1 + sA_0RC)(1 + s/\omega_{ta})} \tag{12.153}$$

A bode plot of $H(s)$ appears in Fig. 12.42. Rather than the infinite dc gain provided by an ideal integrator, the dc gain is simply that of the opamp, A_0. At angular frequencies above $1/A_0RC$, the frequency response is a good approximation of the ideal integrator, $-1/sRC$. Then, at frequencies around the unity gain frequency of the opamp, the integrator again deviates from its ideal value due to the limited gain-bandwidth of the opamp. Hence, the integrator is well behaved over the frequency range $1/A_0RC < \omega < \omega_{ta}$ which may place requirements upon the opamp's dc gain and unity gain frequency.

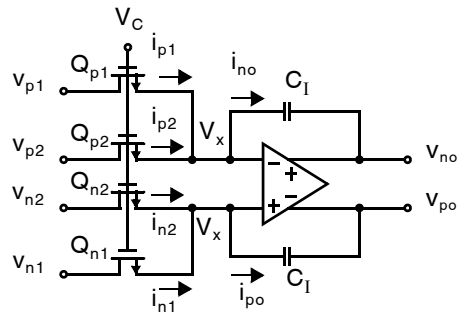


Fig. 12.43 A two-input two-transistor MOSFET-C integrator. The MOS transistors are in the triode region.

Based on the block diagram shown in Fig. 12.2, an example active RC implementation of the two-integrator biquad is illustrated in Fig. 12.44(*a*). The negative conductances are easily realized by cross-connections in a fully-differential implementation. Note that two opamps are required: one for each integrator—hence, one for each pole in the filter. The output of the first integrator is given by

$$v_1(s) = -\frac{G_1}{sC_A}v_i(s) - \frac{G_4}{sC_A}v_o(s) \tag{12.154}$$

whereas the second integrator's output is given by

$$sC_B v_o(s) = G_3 v_1(s) - G_2 v_i(s) - sC_1 v_i(s) - G_5 v_o(s) \tag{12.155}$$

Combining the preceding two equations and rearranging, we have

$$\frac{v_o(s)}{v_i(s)} = -\frac{\left[\left(\frac{C_1}{C_B}\right)s^2 + \left(\frac{G_2}{C_B}\right)s + \left(\frac{G_1G_3}{C_AC_B}\right)\right]}{\left[s^2 + \left(\frac{G_5}{C_B}\right)s + \left(\frac{G_3G_4}{C_AC_B}\right)\right]} \tag{12.156}$$

Thus, we see that the transfer function is general, and that negative coefficients can be realized by cross-coupling appropriate signal wires.

12.8.2 MOSFET-C Two-Transistor Integrators

MOSFET-C filters are similar to fully differential active-RC filters, except resistors are replaced with equivalent MOS transistors in triode. Since active-RC and MOSFET-C filters are closely related, one benefit of MOSFET-C filters is that a designer can use the wealth of knowledge available for active-RC filters. Here, we discuss three variations of MOSFET-C filters—two-transistor integrators, four-transistor integrators, and R-MOSFET-C filters.

An example of a two-transistor MOSFET-C integrator is shown in Fig. 12.43 [Banu, 1983]. It is the counterpart of the two-input fully differential active-RC integrator of Fig. 12.41(*b*). Before we look at the linearization mechanism that occurs in a MOSFET-C integrator, let us look at the equivalence of the two circuits in Fig. 12.41(*b*) and Fig. 12.43. In the active-RC filter, the two input voltages of the opamp are equal (due to negative feedback) and, defining these two input voltages to be V_x, we can write

$$i_{no} = i_{p1} + i_{p2} = \frac{v_{p1} - V_x}{R_{p1}} + \frac{v_{p2} - V_x}{R_{p2}} \tag{12.157}$$

and

$$i_{po} = i_{n1} + i_{n2} = \frac{v_{n1} - V_x}{R_{n1}} + \frac{v_{n2} - V_x}{R_{n2}} \tag{12.158}$$

The output signals are given by

$$v_{po} = V_x - \frac{i_{po}}{sC_I} \tag{12.159}$$

and

$$v_{no} = V_x - \frac{i_{no}}{sC_I} \tag{12.160}$$

Thus, the differential output signal, v_{diff}, is

$$v_{diff} \equiv v_{po} - v_{no} = \frac{i_{no} - i_{po}}{sC_I} \tag{12.161}$$

Assuming the component values in the positive and negative half circuits are equal, or specifically, if

$$R_1 \equiv R_{p1} = R_{n1} \quad \text{and} \quad R_2 \equiv R_{p2} = R_{n2} \tag{12.162}$$

we have

$$v_{diff} = \frac{1}{sR_1C_I}(v_{p1} - v_{n1}) + \frac{1}{sR_2C_I}(v_{p2} - v_{n2}) \tag{12.163}$$

Thus, we see that the differential output signal, v_{diff}, equals positive integration of the two input differential signals. Negative integration is easily obtained by cross-coupling two input or output wires. Note that the common-mode signal, V_x, does not affect the final result.

Since the transistors in the MOSFET-C integrator, shown in Fig. 12.43, are assumed to be in the triode region, the small-signal analysis for the MOSFET-C integrator is quite similar to that for the active-RC circuit. Recall from (12.57) that the small-signal equivalent resistance of a transistor in triode is given by

$$r_{DS} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)\left(V_{GS} - V_{tn}\right)} \tag{12.164}$$

Thus, assuming the transistors are matched, the differential output of the MOSFET-C integrator is given by

$$v_{diff} = \frac{1}{sr_{DS1}C_I}(v_{p1} - v_{n1}) + \frac{1}{sr_{DS2}C_I}(v_{p2} - v_{n2}) \tag{12.165}$$

where

$$r_{DSi} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_i\left(V_C - V_x - V_{tn}\right)} \tag{12.166}$$

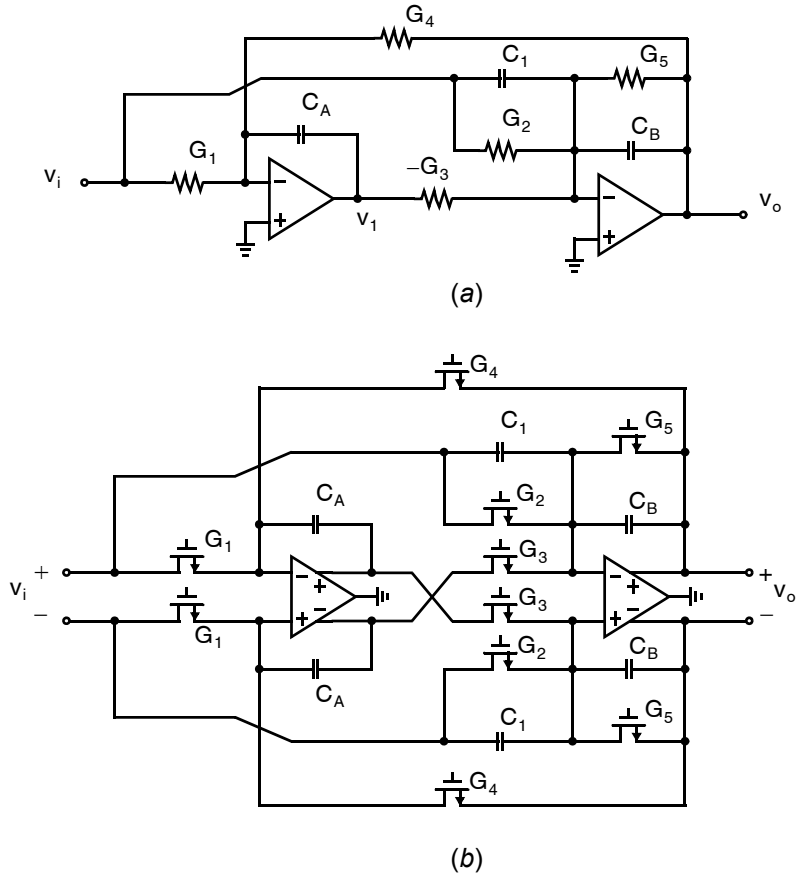


Fig. 12.44 (*a*) A general second-order active-RC half circuit. (*b*) The equivalent MOSFET-C filter.

Here, note that V_x is important because it helps to determine the small-signal resistance of the transistors. Tuning can be accomplished by varying the control voltage, V_c.

Key Point: *Replacing the fixed resistors of an active RC integrator with transistors in triode provides tunability, but reduces linearity.*

So far, we have analyzed the small-signal operation of a MOSFET-C integrator. In this MOSFET-C case, the differential input signals, (v_{p1}, v_{n1}) and (v_{p2}, v_{n2}), are assumed to be balanced around a common-mode voltage, V_{CM}, set by the output common-mode feedback from the previous integrators. The common-mode feedback of the integrator causes the integrator's two outputs to be balanced also. As a result, it can be shown that V_x is related to the square of the input signal. This nonlinear relationship at V_x together with the nonlinear current through triode transistors results in quite linear single-ended signals at V_{no} and V_{po}. In addition, since the circuit is fully differential, even-order distortion products cancel, as depicted in Section 6.7. Thus, two-transistor MOSFET-C integrators can be used to realize filters with around a 50-dB linearity.

An example of a general second-order MOSFET-C filter is shown in Fig. 12.44(*b*). It is equivalent to the active RC filter half-circuit of Fig. 12.44(*a*), with the terms G_i representing the small-signal drain-source conductance of the transistors. The transfer function for this filter is the same as its equivalent active-RC half circuit, given by (12.156).

12.8.3 Four-Transistor Integrators

One way to improve the linearity of a MOSFET-C filter is to use four-transistor MOSFET-C integrators, as shown in Fig. 12.45 [Czarnul, 1986]. The small-signal analysis of this four-transistor integrator can be found by treating this one-input integrator as effectively a two-input integrator in which the two input signals are $(v_{pi} - v_{ni})$ and the inverted signal is $(v_{ni} - v_{pi})$. With such an observation, and applying (12.165), the output signal is given by

$$v_{diff} = \frac{1}{sr_{DS1}C_I}(v_{pi} - v_{ni}) + \frac{1}{sr_{DS2}C_I}(v_{ni} - v_{pi}) \tag{12.167}$$

where

$$r_{DS1} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_1\left(V_{C1} - V_x - V_t\right)} \tag{12.168}$$

and

$$r_{DS2} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_2\left(V_{C2} - V_x - V_t\right)} \tag{12.169}$$

In the case where all four transistors are matched, we can combine the preceding equations to obtain

$$v_{diff} = v_{po} - v_{no} = \frac{1}{sr_{DS}C_I}(v_{pi} - v_{ni}) \tag{12.170}$$

where r_{DS} is given by

$$r_{DS} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)\left(V_{C1} - V_{C2}\right)} \tag{12.171}$$

Thus, we see that the effective resistance is determined by the difference in the control voltages.

With regard to the distortion properties of this integrator, it has been shown for long channel-length transistors that the drain-source current in a triode transistor is reasonably well modelled by an equation in which the nonlinear distortion terms do not depend on the controlling gate voltage. In other words, although the linear term of each drain

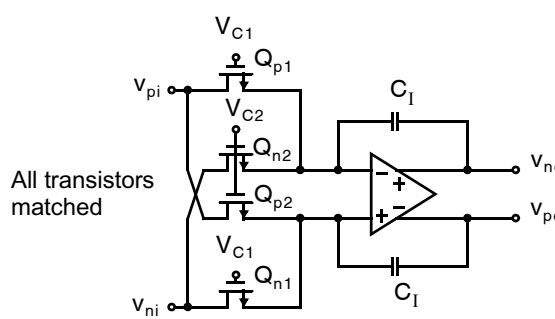


Fig. 12.45 A single-input four-transistor MOSFET-C integrator. It can cancel both even and odd distortion products if the nonlinear terms are not dependent on the control signals, V_{C1} and V_{C2}.

current is set by the controlling gate voltage, all distortion terms of the drain currents remain equal in the pair Q_{p1}, Q_{p2} as well as in the pair, Q_{n1}, Q_{n2} since, within each pair, the drain-source voltages are equal. As a result of the cross-coupling connection, *both the even and odd distortion terms cancel*. Unfortunately, this model does not hold as well for short-channel devices, and transistor mismatch also limits the achievable distortion performance. Hence, in practice, one might expect around a 10-dB linearity improvement using this technique over the two-transistor MOSFET-C integrator.

12.8.4 R-MOSFET-C Filters

Another technique intended to improve the linearity of MOSFET-C filters is the use of additional linear resistors [Moon, 1993]. An R-MOSFET-C first-order filter and its MOSFET-C counterpart are shown in Fig. 12.46. Using the equivalent active-RC half circuit, shown in Fig. 12.46(*c*), one finds the transfer function of this first-order R-MOSFET-C filter to be given by

$$\frac{v_o(s)}{v_i(s)} = -\frac{R_2/R_1}{sC_I\left[R_2\left(1+\dfrac{R_{Q1}}{R_1 \parallel R_2 \parallel R_{Q2}}\right)\right]+1} \tag{12.172}$$

Key Point: *Adding fixed passive series resistors to active MOSFET-C filters can improve their linearity.*

where R_{Q1} and R_{Q2} represent the small-signal drain-source resistances of Q_1 and Q_2, respectively. Note that the dc gain of this first-order filter is not adjustable. However, it can be set precisely since it is determined by the ratio of two resistors. In addition, the time constant can vary from a minimum value of R_2C_I to an arbitrarily large value by changing the values of R_{Q1} and R_{Q2} via the control voltages, V_{C1} and V_{C2}. What this transfer function equation indicates is that at low frequencies, the feedback around R_2 forces node v_x to act as a virtual ground, and therefore excellent linearity is obtained since the circuit is essentially a resistor-only circuit. In [Moon, 1993], a linearity of approximately –90 dB is observed with input signals in the range of a few kilohertz. At higher frequencies, the linearity steadily decreases, and measured results of –70 dB are reported at the passband edge of a fifth-order Bessel filter (the passband is around 20 kHz). However, it should be noted that the measured values are total-harmonic-distortion values, and no intermodulation results are reported.

One other approach that has been used to realize high-linearity continuous-time filters is simply to realize differential active-RC filters, but to realize the capacitors using digitally programmable capacitor arrays. For example, [Durham, 1993] describes a filter intended for digital-audio interfaces that achieves distortion products at a –94-dB and a 95-dB signal-to-noise ratio.

12.9 TUNING CIRCUITRY

Key Point: *Analog filter time constants, determined by* G_m/C *or* RC *products, vary considerably over process and temperature. Hence, tuning is required to ensure accurate and repeatable frequency responses.*

As mentioned at the beginning of this chapter, one disadvantage of continuous-time integrated filters is the requirement of additional tuning circuitry. This requirement is a result of large time-constant fluctuations due mainly to process variations. For example, integrated capacitors might have 10 percent variations, whereas resistors or transconductance values might have around 20 percent variations. Since these elements are constructed quite differently, the resulting RC or G_m/C time-constant products are accurate to only around 30 percent with these process variations. In addition, temperature variations further aggravate the situation because resistance and transconductance values change substantially

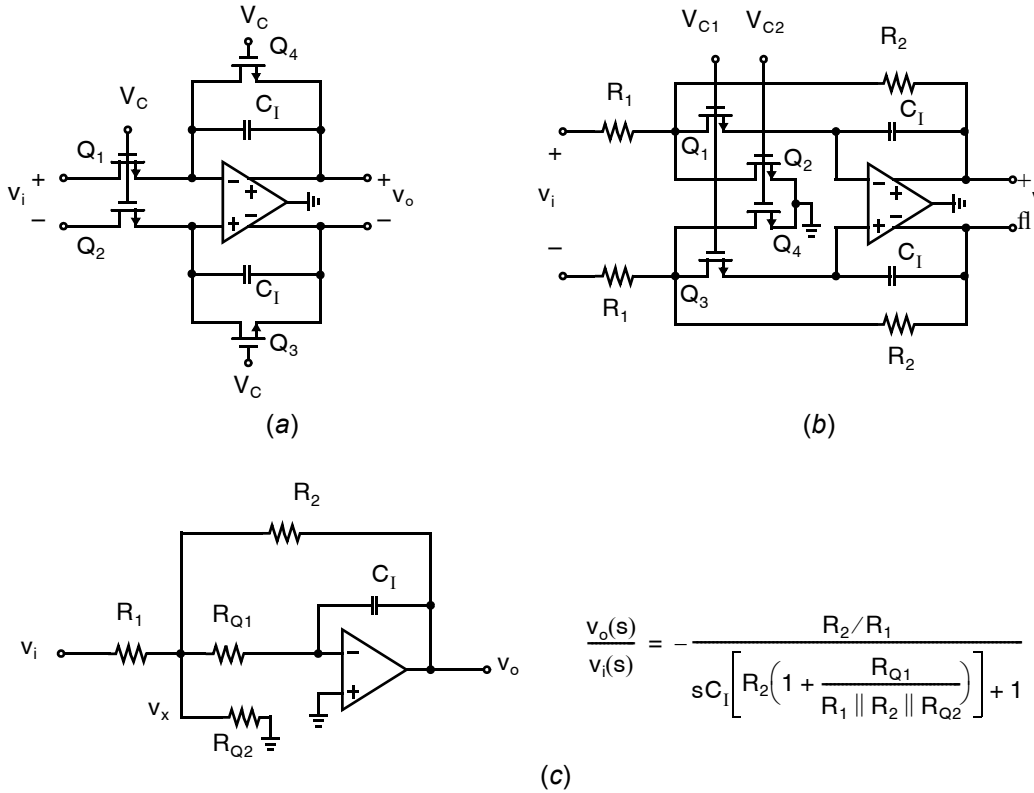


Fig. 12.46 (*a*) A first-order MOSFET-C filter. (*b*) A first-order R-MOSFET-C filter. (*c*) Equivalent active-RC half-circuit for R-MOSFET-C filter.

over temperature. Thus, we need tuning circuitry to modify transconductance values such that the resulting overall time constants are set to known values. It should be noted, however, that although absolute component tolerances are high, relative accuracies can be quite good. Specifically, the ratio of two capacitors (or two resistors, or two transistors) can be matched to under 1 percent. This section discusses some techniques to realize tuning circuitry for integrated continuous-time filters. Note that, although the tuning approaches are discussed while we refer to G_m-C (transconductance-C) filters, the concepts are easily extended to other integrated continuous-time filters, such as MOSFET-C filters.

12.9.1 Tuning Overview

Consider once again the G_m-C continuous-time biquad filter shown in Fig. 12.10. Its transfer function, repeated here for convenience, is given by

$$H(s) = \frac{s^2\left[\frac{CK_X}{C_X+C_B}\right] + s\left[\frac{G_{m5}}{C_X+C_B}\right] + \left(\frac{G_{m2}G_{m4}}{[C_A(C_X+C_B)]}\right)}{s^2 + s\left[\frac{G_{m3}}{C_X+C_B}\right] + \left(\frac{G_{m1}G_{m2}}{[C_A(C_X+C_B)]}\right)} \tag{12.173}$$

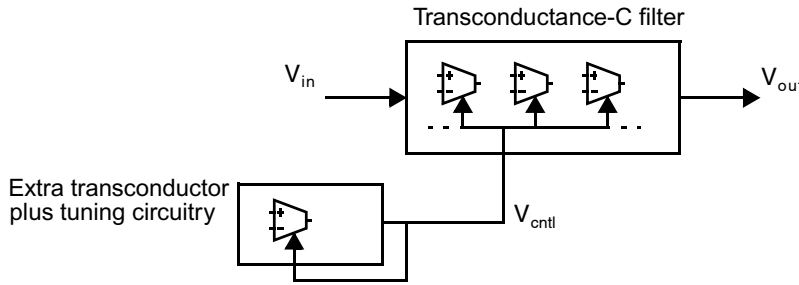


Fig. 12.47 Indirect frequency tuning. Note that, although the tuning signal is shown here as a voltage signal, it may also be realized using a set of current controlling signals.

Although the transconductances, G_{mi}, have large variations in their absolute values, relative transconductance values can also be set reasonably accurately. For example, if the transconductors are realized as bipolar differential pairs with an emitter degeneration resistor (as shown in Fig. 12.12), then the relative transconductance values of two transconductors is set by the ratio of their respective emitter degeneration resistors (assuming the gain-cell currents are equal in the two transconductors). Similarly, if other transconductors of the same type are used, their relative G_m values can be accurately set by choosing appropriate transistor sizes, bias currents, or resistor values (or some combination of the three). As a result, we can write all the transconductance values, G_{mi}, as an accurate constant multiplied by some basic transconductance value. Specifically, let us write all the transconductance values, G_{mi}, as a function of G_{m1}, resulting in

$$G_{mi} = k_{mi} G_{m1} \tag{12.174}$$

Here, the constants, k_{mi}, are set reasonably accurately by appropriate transistor, resistor, or bias-current scaling. Furthermore, let us also write

$$C_X + C_B = k_{XB} C_A \tag{12.175}$$

where k_{XB} is a constant that is accurately set as the ratio of capacitance values. Similarly, let $k_X = C_X/C_A$. With such an approach, (12.173) becomes

$$H(s) = \frac{s^2\left[\frac{k_X}{k_{XB}}\right] + s\left(\frac{k_{m5}}{k_{XB}}\right)\left(\frac{G_{m1}}{C_A}\right) + \left(\frac{k_{m2}k_{m4}}{k_{XB}}\right)\left(\frac{G_{m1}}{C_A}\right)^2}{s^2 + s\left(\frac{k_{m3}}{k_{XB}}\right)\left(\frac{G_{m1}}{C_A}\right) + \left(\frac{k_{m1}k_{m2}}{k_{XB}}\right)\left(\frac{G_{m1}}{C_A}\right)^2} \tag{12.176}$$

We see from (12.176) that the coefficients for this second-order filter can be accurately set if only one ratio is tuned—that of G_{m1}/C_A. Since it is difficult to tune C_A, the ratio G_{m1}/C_A is tuned by changing G_{m1}. Of course, as G_{m1} is varied, all the other transconductance values, G_{mi}, are also changed since their values must track that of G_{m1} by their respective constants, k_{mi}. It should be mentioned here that this same single transconductance tuning is also valid for higher-order filter structures.

In summary, the most common way to tune a continuous-time integrated filter is to build an extra transconductor that is tuned (tuning details are discussed next) and to use the resulting tuning signal to control the filter transconductors, as shown in Fig. 12.47. Such an approach is commonly referred to as *indirect frequency tuning*. This tuning is indirect since one relies on matching between the filter transconductors (and parasitic capacitances) and the extra transconductor. In other words, the filter is not directly tuned by looking at its output signal. Frequency tuning refers to the fact that the ratio G_{m1}/C_A is equal to the unity-gain frequency of a G_m-C integrator. (G_{m1}/C_A is also proportional to the pole frequency, ω_o, in a first- or second-order filter.)

Unfortunately, this frequency-tuning approach may not be sufficient in some applications due to nonideal integrators or mismatch effects. In these applications, some extra tuning circuitry known as Q-factor tuning (discussed later in this section) can be used. It is also possible to implement *direct tuning* where the filter is tuned by directly looking at the filter's output. Two such direct approaches are also discussed in this section.

12.9.2 Constant Transconductance

As just discussed, if no tuning is performed, one might expect around a 30-percent tolerance on the absolute value for the G_m/C ratio. However, integrated capacitance tolerance typically accounts for only a 10-percent variation in the overall 30-percent tolerance. Thus, in applications where a 10-percent variation can be permitted (such as an anti-aliasing filter with a sample rate moderately higher than the Nyquist rate), it is sufficient to set only the G_m value with the use of a fixed external resistor. As we will see below, such tuning of a G_m value is not difficult and does not require any clock signals.

In the case where 10-percent accuracy is not enough, clearly one must also account for capacitor tolerance. One way to do so (while still not needing a clock signal) is to have a variety of external resistors available and choose the most appropriate one to also account for capacitor tolerance. With such an approach, one might expect that the overall G_m/C ratio would be accurate to under 1 percent. Although a variety of methods exist for choosing this external resistance, one way is to use a known value of external resistance and measure the step response of the filter during system production. This step-response result can then be used to determine the C value, and hence the proper resistance value can be calculated. Finally, since integrated capacitors typically have small temperature coefficients, overall temperature effects are small if the external resistor is also chosen to have a low temperature coefficient.

To set a transconductance value equal to the inverse of an external resistance, (i.e., $G_m = 1/R_{ext}$), a variety of feedback circuits can be used. Two example circuits are shown in Fig. 12.48, where it should be noted that, in both cases, it is assumed the transconductor's transconductance increases as the level of the control signal is increased. In Fig. 12.48(*a*), if G_m is too small, the current through R_{ext} is larger than the current supplied by the transconductor, and the difference between these two currents is integrated with the opamp and capacitor. As a result, the control voltage, V_{cntl}, is increased until these two currents are equal (but opposite in direction) and, therefore, $G_m = 1/R_{ext}$. The circuit in Fig. 12.48(*b*) shows two voltage-to-current converters—one is the transconductor being tuned, and the other is a fixed voltage-to-current converter that simply needs a large transconductance and is

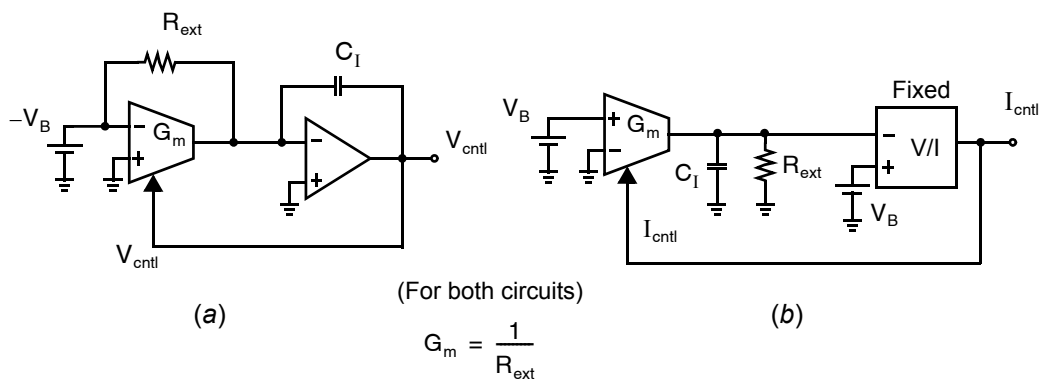


Fig. 12.48 Constant transconductance tuning. Possible tuning circuits if the transconductors are (*a*) voltage controlled or (*b*) current controlled. In both circuits, it is assumed the transconductor's transconductance increases as the level of the control signal increases. V_B is an arbitrary voltage level, whereas C_I is an integrating capacitor used to maintain loop stability.

not necessarily linear. (It might be implemented as a simple differential pair.) This circuit operates as follows: If G_m is too small, then the voltage at the top of R_{ext} will be less than V_B and the fixed V/I converter will increase I_{cntl}. At steady state, the differential voltage into the fixed V/I converter will be zero, resulting in $G_m = 1/R_{ext}$.

12.9.3 Frequency Tuning

A precise tuning of a G_m/C_A ratio can be achieved if an accurate time period is available. For example, assuming one has available an accurate clock frequency, say, f_{clk}, then one method of frequency tuning is by using a switch-capacitor circuit as shown in Fig. 12.49 [Viswanathan, 1982]. Here, the tuning circuit is quite similar to the constant-transconductance approaches of Fig. 12.48, except that the external resistance is replaced with a switched-capacitor equivalent resistance. From Chapter 14, the equivalence resistance of the switched-capacitor circuit is given by $R_{eq} = 1/(f_{clk}C_m)$, and therefore the transconductance value, G_m, is set to $f_{clk}C_m$. In this way, the G_m/C_A ratio is accurately set to $(f_{clk}C_m)/C_A$, and hence precise frequency tuning can be achieved without any external components. Note that an additional low-pass filter, R_1C_1, has been added here to remove the high-frequency ripple voltage due to using a switched-capacitor circuit. It should also be noted that, since this method requires a clock signal, there is a strong possibility that some of the clock signal will leak into the continuous-time filter, either mainly through the transconductor controlling signal or through the IC substrate.

One difficulty with this switched-capacitor tuning approach is that it requires large transconductance ratios (and hence poorer matching) if high-frequency filters are desired. For example, consider the case of a 100-MHz filter, where the G_m/C_A ratio should be set to around $2\pi \times 100$ MHz. If a capacitance value, C_A, is assumed to be 1 pF, then $G_m = 2\pi \times 10^{-4}$ V/A. Thus, in the tuning circuit of Fig. 12.49, assuming that $C_m = 1$ pF, we would require the switched-capacitor circuit to be clocked at the impractical frequency of

$$f_{clk} = \frac{G_m}{C_m} = \frac{2\pi \times 10^{-4}}{1 \times 10^{-12}} = 628 \text{ MHz} \tag{12.177}$$

At first glance, one might consider reducing this clock frequency by increasing the value of C_m, but, although f_{clk} is reduced, settling-time requirements remain difficult since the capacitance, C_m, is larger. Another way to reduce f_{clk} is to let the tuning circuit set a smaller transconductance value, say, $0.1G_m$. In this way, f_{clk} is reduced by ten but the filter's transconductors must be 10 times greater than the tuning circuitry's transconductor. Thus, poorer matching occurs between the filter and the tuning circuitry, resulting in a less accurate frequency setting for the filter.

Another approach to lower the clock frequency of this switched-capacitor tuning circuitry is by using two scaled current sources, as shown in Fig. 12.50 [SilvaMartinez, 1992]. Using an equivalent resistance of

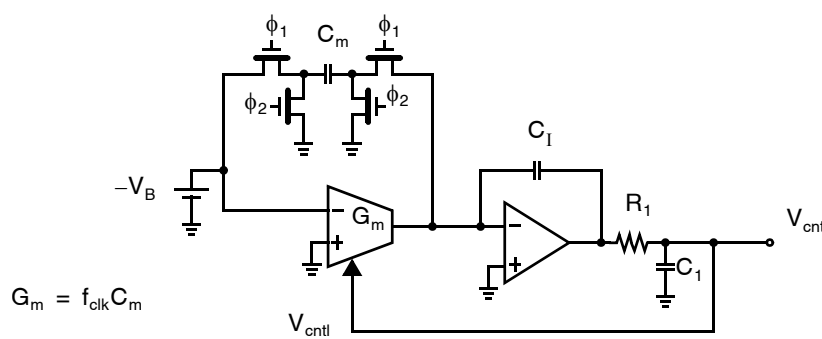


Fig. 12.49 A frequency-tuning circuit where G_m is set by a switched-capacitor circuit. The switched-capacitor branch has a clock frequency of f_{clk}. The low-pass filter, R_1C_1, is set to a frequency low enough to reduce the high-frequency ripple voltage.

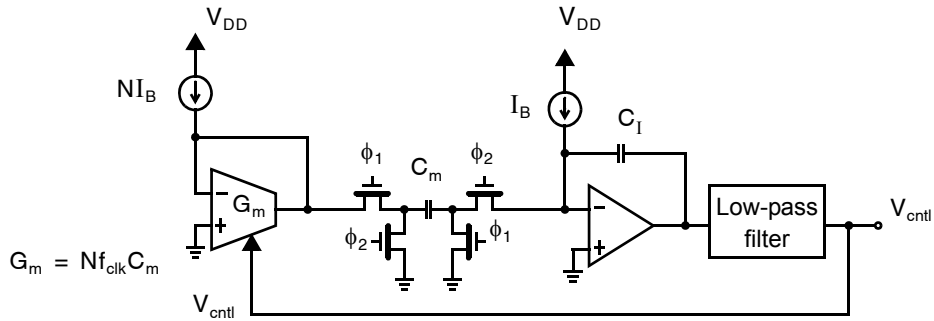


Fig. 12.50 A switched-capacitor frequency-tuning circuit that operates at a lower clock frequency.

$-1/f_{clk}C_m$ for the switched-capacitance circuit, and noting that the diode-connected transconductor is equivalent to a resistor of value $1/G_m$, one can show that when the average current into the integrator is zero, the transconductance value is given by

$$G_m = Nf_{clk}C_m \tag{12.178}$$

Thus, the clock frequency of this circuit is N times lower than that for the circuit shown in 15.47. This circuit was used to tune a 10.7-MHz bandpass filter, where a value of $N = 148$ was used, resulting in a clock frequency of only 450 kHz. The low-pass filter was also realized by using a switched-capacitor circuit and an off-chip capacitor.

Another approach to achieving frequency tuning is to use a phase-locked loop (PLL) (described in Chapter 19) as shown in Fig. 12.51 [Tan, 1978]. Here, two continuous-time integrators are placed in a loop to realize a voltage-controlled oscillator that is placed into a phase-locked loop. After the circuit is powered up, the negative feedback of the PLL causes the VCO frequency and phase to lock to the external reference clock. Once the VCO output is locked to an external reference signal, the G_m/C ratio of the VCO is set to a desired value, and the control voltage, V_{cntl}, can be used to tune the integrated filter. It should be noted that choosing the external reference clock is a trade-off because it affects both the tuning accuracy as well as the tuning signal leak into the main filter. Specifically, for best matching between the tuning circuitry and the main filter, it is best to choose the reference frequency that is equal to the filter's upper passband edge. However, noise gains for the main filter are typically the largest at the filter's upper passband edge, and therefore the reference-signal leak into the main filter's output might be too severe. As one moves away from the upper passband edge, the matching will be poorer, but an improved immunity to the reference signal results. Another problem with this approach is that, unless some kind of power-supply-insensitive voltage control is added to the VCO, any power-supply noise will inject jitter into the control signal, V_{cntl}. The addition of amplitude control is quite complicated. For a good discussion of this phase-locked-loop approach, see [Krummenacher, 1988; Khoury, 1991]. For a better understanding of the operation of a phase-locked loop, see Chapter 19.

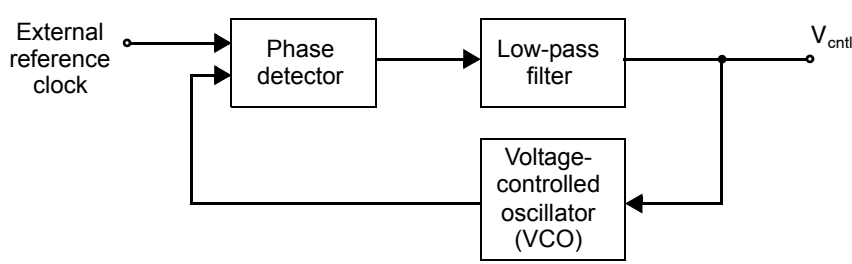


Fig. 12.51 Frequency tuning using a phase-locked loop. The VCO is realized by using transconductor-based integrators that are tuned to adjust the VCO's frequency.

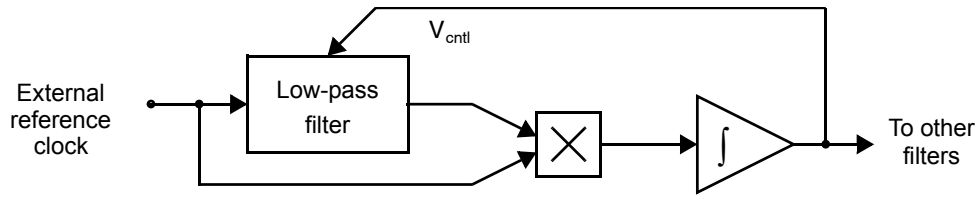


Fig. 12.52 Using a tracking filter to derive a control signal for tuning continuous-time filters.

Another, somewhat similar, approach that has been used is based on having a tracking filter that locks onto a fixed-frequency input [Khorramabadi, 1984]. This approach is reported to be less sensitive to power-supply noise. The control voltage for the tracking filter is used to adjust the transconductors that are used in the actual system, assuming the transconductors in the tracking filter match the transconductors in the actual system. An example calibration system is shown in Fig. 12.52. When the low-pass filter has the same resonant frequency as the external reference clock, it has exactly a 90-degree phase shift. Therefore, the correlation between its input and output is zero. If the frequencies are not the same, then the correlation, as obtained by the multiplier and integrator, is nonzero. The negative feedback from the output of the integrator to the frequency-control voltage of the filter drives the resonant frequency in the proper direction to equalize the two frequencies. After lock is obtained, the same control voltage can be used for other filters. The other filters will be tuned to known time constants, assuming they match the filter in the control loop.

It is important that the integrator in the loop have a very low-frequency time constant in order to filter out the component at twice the external reference frequency that is present at the output of the multiplier. Otherwise, this component can cause clock leakage into the other filters being tuned. A variant on this scheme drives the correlation between a notch output and a low-pass filter output (obtained from the same biquad) to zero when one is obtaining the control voltage [Kwan, 1991]. This latter method has much fewer signal components at twice the frequency of the external reference clock and does not require as large a time constant for the integrator in the loop. This makes the latter method more readily integrable without external components. Also, the loop is less sensitive to phase errors; it will always converge to the notch frequency, even when the difference in phase between the low-pass and notch outputs is not exactly $90°$.

12.9.4 Q-Factor Tuning

In some applications, such as high-speed or highly-selective filtering, nonideal integrator effects and parasitic components force the need to also tune the Q-factors of the poles of the integrated filter. Although we have seen that tuning a single time constant can set all coefficients of an integrated filter with around 1% accuracy, when $Q > 1$, even small errors in the Q-factor can result in large errors in the filter frequency and step response. Hence, in some applications additional tuning is needed to precisely set the Q-factors.

One way to perform Q-factor tuning is to tune the phase of the filter's integrators to ensure that they all have a 90-degree phase lag near the filter's passband edge. An integrator's phase response can be adjusted by introducing a tunable resistance (say, realized as a transistor in triode) in series with the integrating capacitor. The control voltage for this tunable resistance, V_Q, is generated by a Q-factor tuning circuit, shown in Fig. 12.53, and passed to all the filter integrators.

Alternatively, in the case where the integrated filter is second order, the pole Q can be adjusted by changing the transconductance of the damping transconductor in the filter. Specifically, referring back to (12.176) and equating the denominator to a second-order denominator written in the form of ω_o and Q, we have

$$s^2 + s\left(\frac{k_{m3}}{k_{XB}}\right)\left(\frac{G_{m1}}{C_A}\right) + \left(\frac{k_{m1}k_{m2}}{k_{XB}}\right)\left(\frac{G_{m1}}{C_A}\right)^2 = s^2 + s\left(\frac{\omega_o}{Q}\right) + \omega_o^2 \tag{12.179}$$

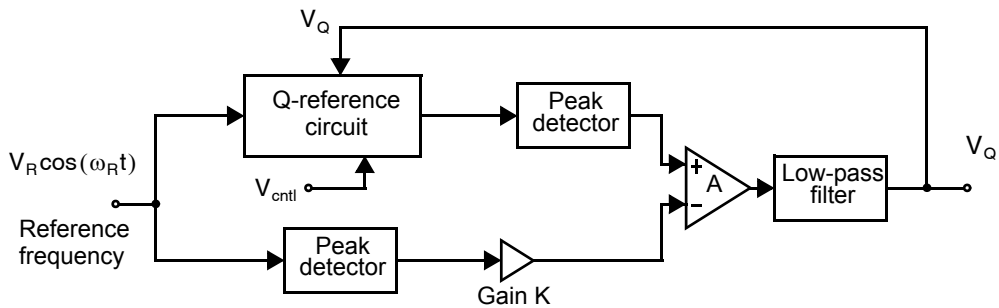


Fig. 12.53 A magnitude-locked loop for Q-factor tuning. The Q-reference circuit is typically a second-order filter that is frequency tuned through the use of V_{cntl}.

Solving, we find the following relationships:

$$\omega_o = \left(\frac{G_{m1}}{C_A}\right)\sqrt{\frac{k_{m1}k_{m2}}{k_{XB}}} \tag{12.180}$$

$$Q = \left(\frac{k_{XB}}{k_{m3}}\right)\sqrt{\frac{k_{m1}k_{m2}}{k_{XB}}} \tag{12.181}$$

From this Q relationship, we see that the Q factor of the filter's poles is inversely proportional to k_{m3}, and hence one need only tune G_{m3} to adjust the Q factor of this second-order filter.

The control signal, V_Q, for either of the preceding Q-factor tuning approaches can be realized using the magnitude-locked loop system shown in Fig. 12.53 [Schaumann, 1990]. The circuit's operation is based on the knowledge that a Q-factor error results in a magnitude error in a filter's frequency response. To measure the magnitude error, a sinusoidal signal of an appropriate reference frequency is applied to a Q-reference circuit, and the magnitude of its output is compared to a scaled version of the input signal's magnitude. This sinusoidal frequency might be obtained as the output of the VCO in the frequency-tuning circuitry.

12.9.5 Tuning Methods Based on Adaptive Filtering

Adaptive filters are commonly used in digital-signal-processing applications such as model matching, channel equalization, and noise (or echo) cancellation. It is also possible to use adaptive filter techniques to tune a continuous-time integrated filter to a given specification. One example of such an approach is shown in Fig. 12.54 [Kozma, 1991], where the adaptive tuning circuitry is used to minimize the error signal. During tuning, the input to the tunable filter is a predetermined output from a pseudorandom (PN) 4-bit sequence and a D/A converter. Since the input is known and repetitive, the ideal output from a desired transfer-function can be precalculated (or simulated) and used as a reference signal for the adaptive system. After adaptation, the transfer function of the tunable filter should match that of the desired filter, and, by freezing the coefficients (assuming they are digitally controlled), the tunable filter can be disconnected from the tuning circuit and used in the desired operation. With this approach, it is possible to tune all the poles and zeros of the tunable filter to better match the desired transfer function.

Another adaptive-filter tuning method intended for high-speed data transmission over a twisted-wire pair is shown in Fig. 12.55 [Shoval, 1995]. In this application, the filter input is a data stream of ones and zeros that are represented by the input signal being either high or low. Before transmission, a pulse-shaping filter is required to ensure that not too much high-frequency power is radiated from the twisted-wire channel. The two tuning loops

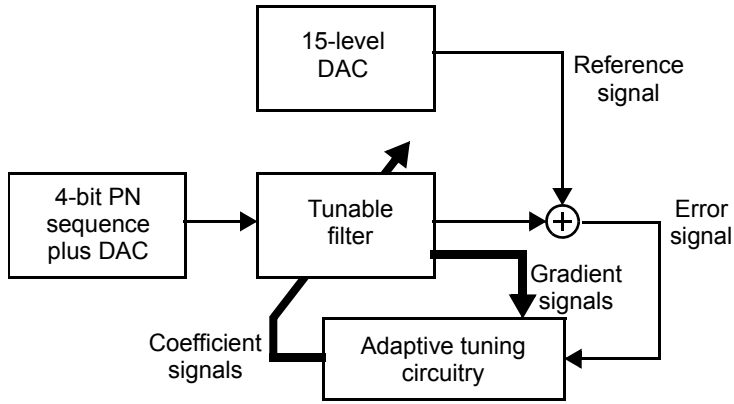


Fig. 12.54 A tuning approach in which all the poles and zeros of the tunable filter can be adjusted. The 15-level DAC outputs are determined by what the ideal filter response would generate, while the adaptive tuning circuitry minimizes the error signal.

(frequency and Q-factor loops) rely on the fact that the input to the filter is a series of steps (say, ± 1 V), and thus, in fact, tune the step response of the filter. Specifically, frequency tuning is accomplished by ensuring that the zero crossing of the filter's low-pass output occurs at the correct time period after a data transition. Also, Q-factor tuning is performed by comparing the peak in the filter's bandpass output with a known voltage. Fortunately, the peak in the bandpass output occurs at approximately the same time as the zero crossing, so both detectors can be realized through the use of two clocked comparators, which are triggered at a set time after a data transition. In summary, clever tuning techniques can be implemented when a restricted set of input signals are applied to a tunable filter such as in data communication circuits.

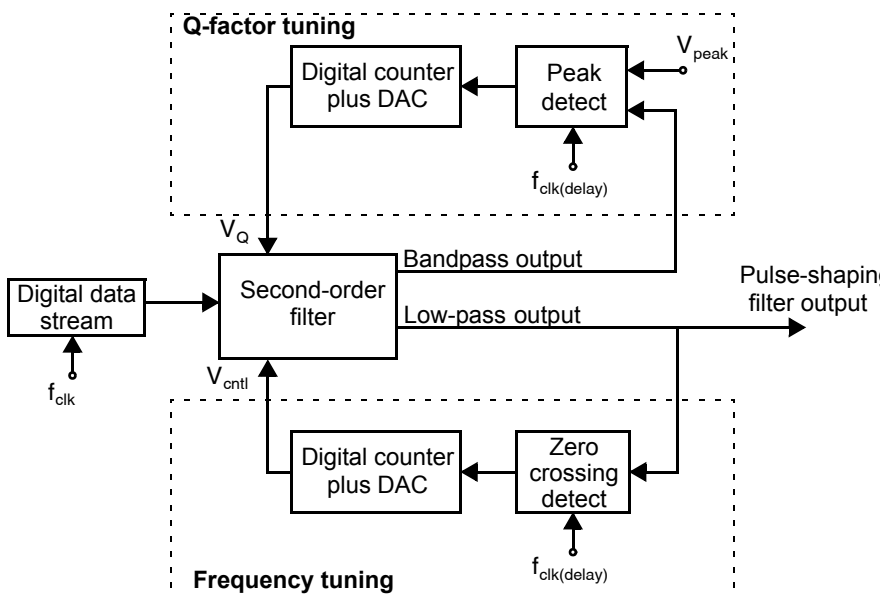


Fig. 12.55 A tuning method to implement a pulse-shaping filter needed for data communication systems.

12.10 INTRODUCTION TO COMPLEX FILTERS

This section is based upon the tutorial introduction published in [Martin, 2004] much of which is in turn based upon work published in [Lang, 1981], [Snelgrove, 1981], [Snelgrove, 1982], [Allen, 1985].

12.10.1 Complex Signal Processing

A *complex* signal consists of a pair of *real* signals at an instant in time. If one denotes the complex signal, $X(t)$, as $X(t) = x_r(t) + jx_q(t)$, where $j = \sqrt{-1}$, then a Hilbert Space can be defined using appropriate definitions for addition, multiplication, and an inner-product and norm. In an actual physical system, the signals are both real (but are called the real and imaginary parts) and are found in two distinct signal paths. The multiplier "j" is used to help define operations between different complex signals, and is to some degree *imaginary*; i.e., it doesn't actually exist in real systems. Often the dependence of signals on time is not shown explicitly.

The use of *complex* signal processing to describe wireless systems is increasingly important and ubiquitous for the following reasons: it often allows for image-reject architectures to be described more compactly and simply; it leads to a graphical or *signal-flow-graph* (SFG) description of signal-processing systems providing insight, and it often leads to the development of new systems where the use of high-frequency highly selective image-reject filters is minimized. The result is more highly integrated systems using less power and requiring less physical space.

As an example, consider the popular "*quadrature*" mixer with a *real* input as shown in Fig. 12.56. In the complex SFG, two real multiplications have been replaced by a single complex multiplication. Furthermore, since in the time domain we have

$$\begin{aligned} Y(t) &= y_r(t) + jy_q(t) \\ &= X(t)e^{j\omega_i t} \end{aligned} \tag{12.182}$$

then taking the Fourier transform we have

$$Y(\omega) = X(\omega - \omega_i) \tag{12.183}$$

and we see that an ideal quadrature mixer results in a simple frequency-shift of the input signal with no images occurring, a conclusion that is obvious from Fig. 12.56(*b*), but is certainly not obvious from Fig. 12.56(*a*). If this mixer is followed by a low-pass filter in each path, then it can be used to directly demodulate carrier-frequency signals without images. For this reason, it often called an *image-reject mixer* although if one is being accurate, the term should denote the combination of the quadrature mixer and lowpass filters.

Complex signal processing systems operate on ordered pairs of real signals. A possible definition for a complex inner-product is

$$\langle x_1(t), x_2(t) \rangle = \lim_{t \to \infty} \frac{1}{2t} \int_{-t}^{t} x_1(\tau) x_2^*(\tau) d\tau \tag{12.184}$$

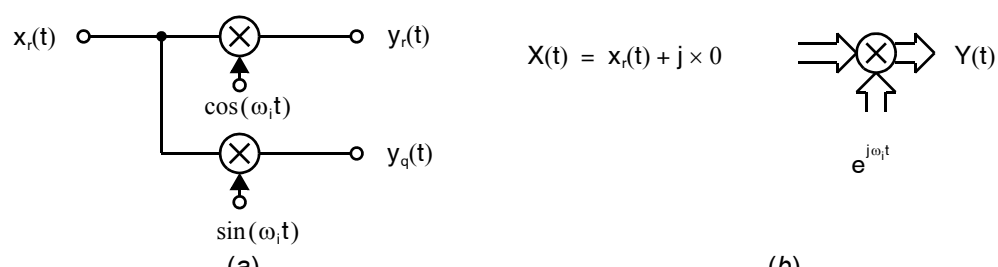


Fig. 12.56 A signal-flow-graph (SFG) of an quadrature mixer using (*a*) a real SFG and (*b*) a complex SFG.

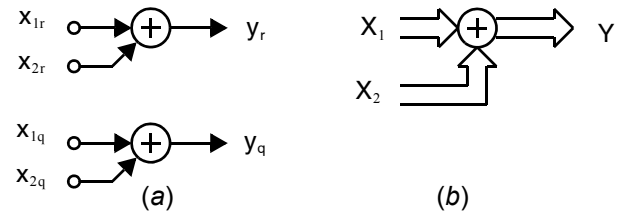


Fig. 12.57 (*a*) The real SFG for adding two complex signals and (*b*) the equivalent complex SFG.

Combining this definition with the typical definitions for complex inversion, addition, and multiplication, and with the properties of commutation and association, constitutes a Hilbert space. It is assumed that all signals are periodic and frequency-limited and therefore can be expressed as a finite series of "*complexoids*"[3]

$$x_i(t) = \sum_{i=-N}^{N} k_i e^{j\omega_0 it} \tag{12.185}$$

where k_i is in general complex. Here, we will limit our consideration to a single (or a few) complexoids in order to simplify explanations.

12.10.2 Complex Operations

Complex continuous-time filters are realized using only three basic operations: complex addition, complex multiplication, and complex integration.

Key Point: *Complex filters are realized using three basic operations performed on ordered pairs of real signals: complex addition, multiplication, and integration.*

A complex addition of two complex signals simply adds the two real parts and the two imaginary parts independently. Consider the real SFG that represents adding a complex signal to a second complex signal, as is shown in Fig. 12.57(*a*). Its complex SFG is shown in Fig. 12.57(*b*).

Slightly more complicated is a complex multiplication; consider multiplying a complexoid $X(t) = e^{j\omega_i t}$ by a complex coefficient $K = a + jb$. We have

$$\begin{aligned} Y(t) = {} & \{a\cos(\omega_i t) - b\sin(\omega_i t)\} \\ & + j\{a\sin(\omega_i t) + b\cos(\omega_i t)\} \end{aligned} \tag{12.186}$$

The real SFG of this operation is shown in Fig. 12.58(*a*). Once again the complex SFG of a multiplication operation, as shown in Fig. 12.58(*b*), is considerably simpler. Note that the real SFG does not include the constant j anywhere, whereas, in the complex SFG, it is admissible.

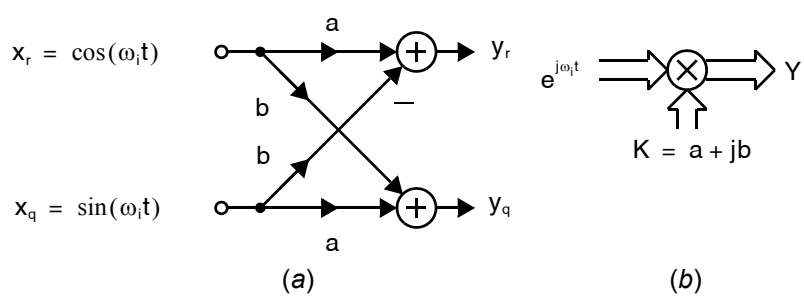


Fig. 12.58 (*a*) The real SFG of a complex multiplication and (*b*) the equivalent complex SFG.

3. Here, the word *"complexoid"* is used to represent the complex single-frequency signal component $e^{j\omega t} = \cos(\omega t) + j\sin(\omega t)$.

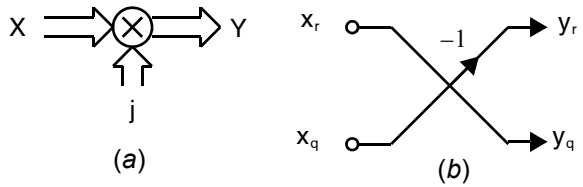


Fig. 12.5 9 Multiplying a signal by the imaginary constant j.

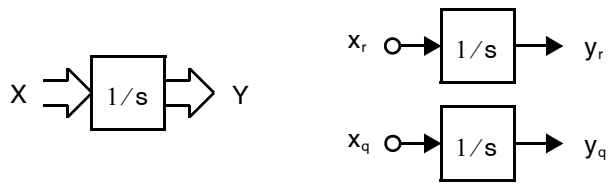


Fig. 12.60 The analog integration operator for complex signals.

A special case is the multiplication of an input signal by the constant j. This is shown using a complex SFG in Fig. 12.59(*a*) and a real SFG in Fig. 12.59(*b*). This operation is simply an interchanging of the real and inverted imaginary parts; this operation is the basis of realizing imaginary components in complex networks; that is, imaginary gains are realized by cross-coupling in the two real networks.

One additional operation is required for realizing complex filters: the integration operator. This simply consists of applying the corresponding real operation to each part of the complex signal, as shown in Fig. 12.60.

12.10.3 Complex Filters

Complex filters are important and widely used signal processing blocks in modern wireless transceivers. Complex filters use cross-coupling between the real and imaginary signal paths in order to realize asymmetrical (in the frequency domain) filters having transfer functions that do not have the conjugate symmetry of real filters. This implies that their transfer functions have complex coefficients. Complex filters can be realized using the basic operations of addition and multiplication and the integrator operator.

Key Point: *Complex filters use cross-coupling between real and imaginary signal paths to realize transfer functions that do not have the frequency-domain conjugate symmetry of real filters.*

A complex transfer function can be realized by using four real SFG filters as shown in Figure 12.61 where we have

$$R(s) = \frac{H(s) + H^*(s)}{2} \tag{12.187}$$

and

$$jQ(s) = \frac{H(s) - H^*(s)}{2} \tag{12.188}$$

When the filters leading from the real and imaginary branches of Fig. 12.61 do not exactly match, imaging errors (plus less harmful magnitude response errors) occur.

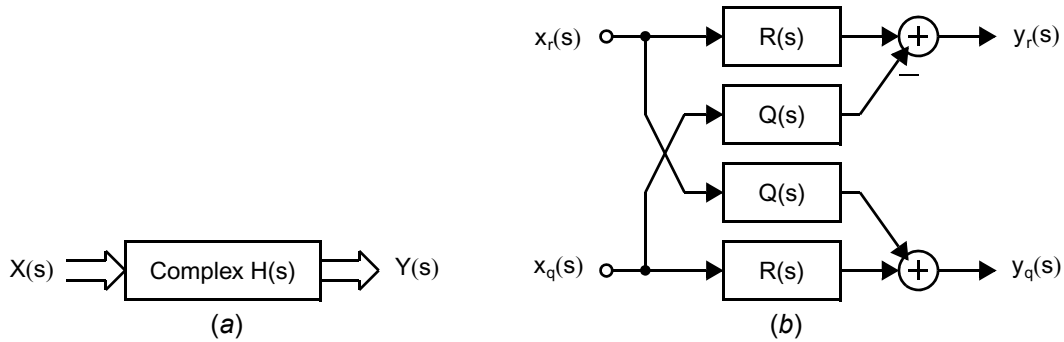


Fig. 12.61 Realizing a complex filter $H(s) = R(s) + jQ(s)$ by real SFG transfer filters $R(s)$ and $Q(s)$.

12.10.4 Frequency-Translated Analog Filters

Key Point: *Real filters can be frequency-translated into complex transfer functions by substituting $(s - j\omega)$ for s.*

Complex analog transfer functions can be realized by frequency-translating real filters [Snelgrove, 1982], [Sedra, 1985]. Specifically, one method for deriving complex transfer functions is to employ a complex substitution for the Laplace variable s. For example, to effect a positive frequency shift, one uses the substitution

$$s \rightarrow s - j\omega_0 \tag{12.189}$$

The effect of this substitution on a low-pass filter is shown in Fig. 12.62

A method for realizing complex analog filters is to start with an active RC prototype filter and then transform it to a complex frequency-translated filter; the procedure to do this is simple:

- First, use an active-RC prototype filter that is fully-differential. This is used to make the realization of negative components trivial.
- Secondly, replicate exactly the prototype filter in order to realize the signal path for the imaginary signal components.
- Finally, add cross-coupled resistors for each capacitor. The size and signs of these resistors determine the frequency translation.

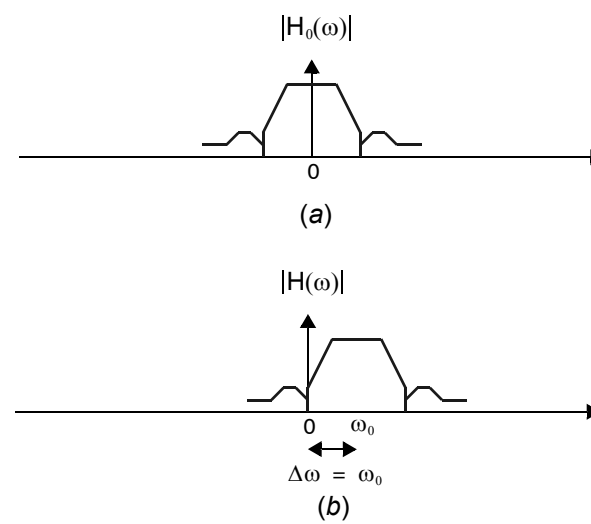


Fig. 12.62 (a) A bas e-band prototype magnitude filter response, and (b) the magnitude r esponse of a complex filt er obtained by the substitution $s \rightarrow s - j\omega_0$.

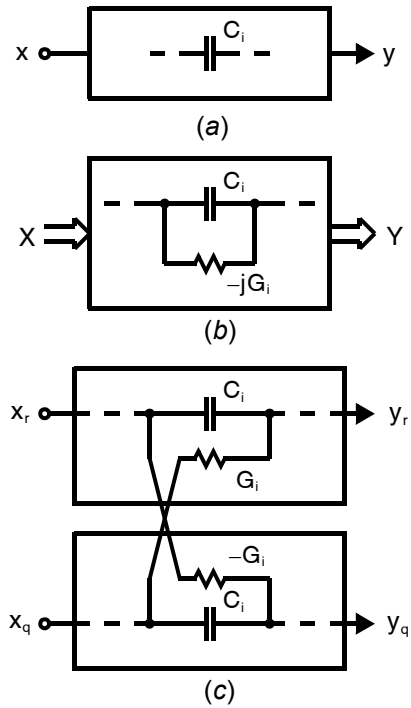


Fig. 12.63 (*a*) A capacitor in the real prototype filter, (*b*) the transformed capacitor in the complex filter, and (*c*) the real-component realization of the complex capacitor.

The second step transforms the real prototype filter into a complex conjugate symmetric filter at the same frequencies as the prototype filter. To see how the third steps results in the desired frequency translation, consider Fig. 12.63.

Fig. 12.63(*a*) shows one of the capacitors in a complex filter at the same frequencies as the real prototype filter. Applying the frequency translation $s \rightarrow s - j\omega_0$ results in transforming the capacitor to an admittance of a capacitor (of the same size) in parallel with an imaginary resistor of size $-j\omega_0 C$ according to the substitution

Key Point: *Beginning with an active-RC prototype filter, duplicating it, and cross-coupling the two signal paths with resistors results in a complex frequency translation of the filter response.*

$$sC \rightarrow sC - j\omega_0 C \tag{12.190}$$

This is shown in Fig. 12.63(*b*) where

$$G_i = \omega_0 C_i \tag{12.191}$$

In order to realize the complex filter using real components, two identical real filters are used with cross-coupling, as is shown in Fig. 12.63(*c*). The negative cross-coupled resistor can be realized using fully differential circuits.

EXAMPLE 12.11

Design a complex filter with a transmission zero at $s = -j\omega_0$.

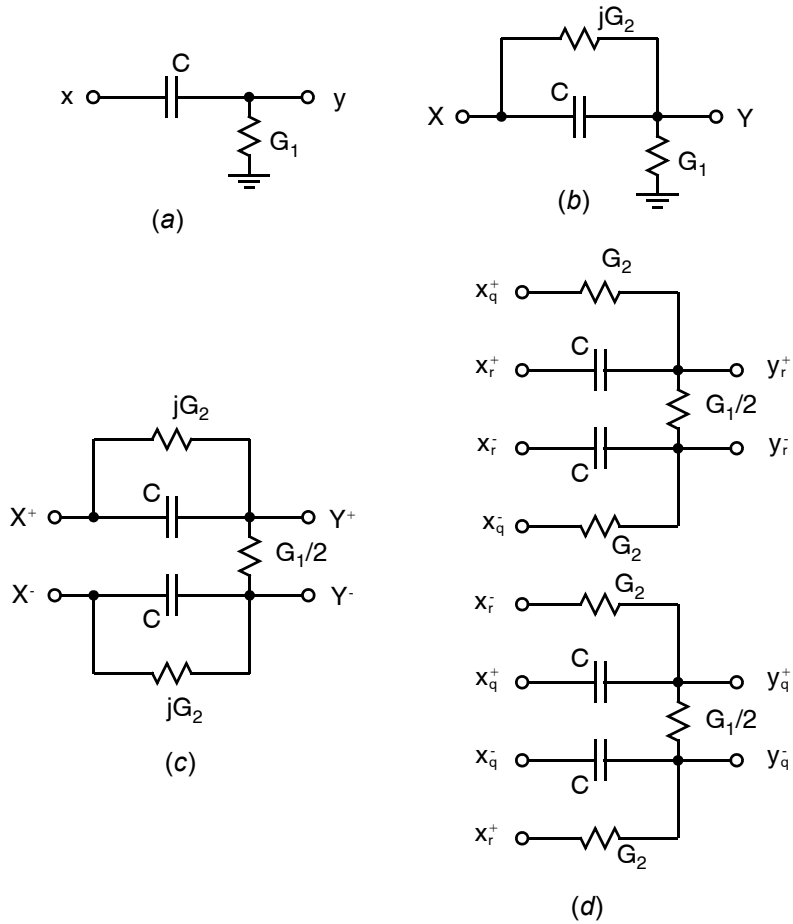


Fig. 12.64 Using a complex frequency translation to realize a PPF: (*a*) the high-pass prototype, (*b*) the complex band-stop having a complex admittance, (*c*) a fully-differential version of (*b*), and (*d*) replacing the complex admittances with cross-coupled real admittances between the real and imaginary paths.

Solution

Begin with the passive RC high-pass filter shown in Fig. 12.64(*a*) which has a transmission zero at $s = 0$. The transmission zero at dc will be frequency-shifted to negative frequencies, so we use the substitution $s \rightarrow s + j\omega_0$. Letting

$$\omega_0 = \frac{G_2}{C} \tag{12.192}$$

gives the required substitution as

$$sC \rightarrow sC + jG_2 \tag{12.193}$$

Thus, the complex frequency-shifted high-pass is now as shown in Fig. 12.64(*b*). A fully-differential version of the circuit of Fig. 12.64(*b*) is shown in Fig. 12.64(*c*). This complex filter can be realized by two parallel real filters

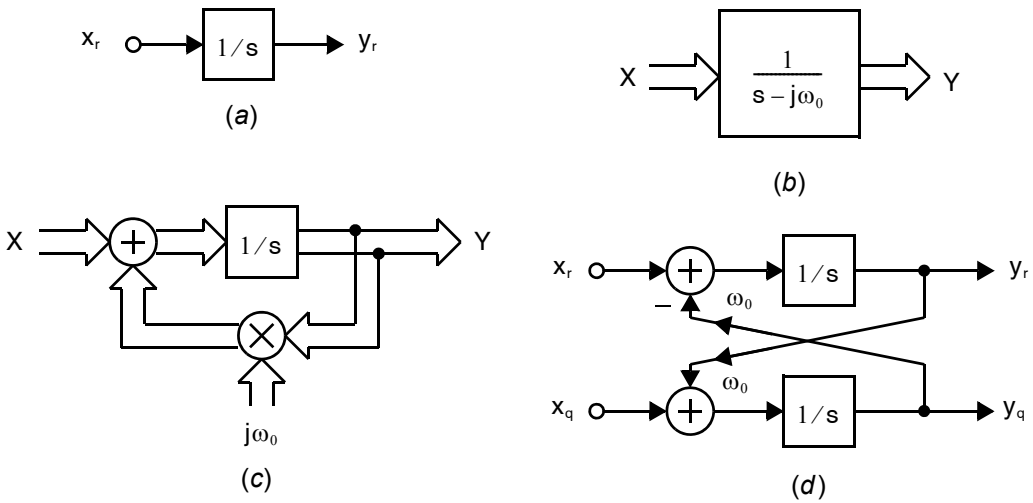


Fig. 12.65 Frequency shifting a real integrator to a complex positive-frequency integrator.

with cross-coupled resistors, as is shown in Fig. 12.64(*d*). This circuit is referred to as a poly-phase network. Taking the load $G_1 = 0$ would result in very narrow notches. Normally, poly-phase networks are intended to be cascaded [Gingell, 1973] and the load on each section is the succeeding section [Behbahani, 2001].

An alternative procedure for realizing continuous-time complex filters is based on frequency-translating SFG prototype filters which are based on integrators [Snelgrove, 1982]. This procedure is illustrated by applying it to the single integrator shown in Fig. 12.65(*a*). In Fig. 12.65(*b*) and (*c*) are shown complex SFGs that realize the frequency-shifted integrator. The complex SFG of Fig. 12.65(*c*) is converted to the equivalent real SFG of Fig. 12.65(*d*), where the necessary cross-coupling is explicit. The procedure for frequency-shifting any real SFG integrator-based filter is simply: a) replicate the real SFG filter for the imaginary path, and b) cross-couple at each integrator, as is shown in Fig. 12.65(*d*); the strength of the cross-couplings determine the frequency of the shift. As always, this cross-coupling is simplified by using fully-differential circuits.

12.11 KEY POINTS

- Continuous-time integrated analog filters process signals that are both continuous in amplitude and time. Discrete-time analog filters process signals sampled in time, and therefore generally operate on lower-frequency signals than their continuous-time counterparts. [p. 469]
- Integrators, summers, and gain stages are the fundamental building blocks of analog filters. [p. 469]
- Any rational transfer function with real-valued coefficients may be factored into first- and second-order terms and, therefore, implemented as a cascade of first- and second-order filters. [p. 470]
- One integrator is required for each pole in an analog filter. [p. 470]
- A popular biquad implementation is two integrators in a loop, one being lossy. The gains around the loop determine its resonant frequency, the integrator loss determines the Q-factor, and the input gains and locations determine the transfer function zeros and gain. [p. 471]

- In G_m-C filters, integrators are realized with a transconductor, which produces a output current proportional to the input voltage, and a capacitor, which integrates the current to produce the output voltage. [p. 472]
- Current signals are readily summed by simply shorting them together. [p. 473]
- Fully-differential circuits are often preferred in analog integrated circuits as they provide improved linearity and noise immunity. [p. 474]
- Differentiating signal paths are sometimes provided by capacitive feed-ins. [p. 475]
- Source-degenerating a differential pair with a resistor linearizes the relationship between differential input voltage and output current. The result is a transconductor with a value approximately equal to the degeneration resistance. [p. 480]
- Feedback may be used to more accurately reproduce the input voltage across the degeneration resistor or further improve the linearity of transconductors with fixed resistors. [p. 483]
- Triode transistors can be used in placed of fixed resistors to make transconductors tunable, although their linearity is generally reduced. [p. 484]
- Transconductors whose gain is based upon the small-signal transconductance of active-mode transistors generally offer increased speed, but reduced linearity, compared with transconductors based upon triode devices. [p. 493]
- Bipolar transconductors based on fixed resistors may be made tunable by cascading them with an adjustable-gain cell. [p. 500]
- Active RC filters employ resistors connected to virtual ground nodes to perform voltage-to-current conversion, and then integrate the resulting currents upon Miller capacitors, connected in feedback around high-gain amplifiers. [p. 510]
- Replacing the fixed resistors of an active RC integrator with transistors in triode provides tunability, but reduces linearity. [p. 514]
- Adding fixed passive series resistors to active MOSFET-C filters can improve their linearity. [p. 516]
- Analog filter time constants, determined by G_m/C or RC products, vary considerably over process and temperature. Hence, tuning is required to ensure accurate and repeatable frequency responses. [p. 516]
- Complex filters are realized using three basic operations performed on ordered pairs of real signals: complex addition, multiplication, and integration. [p. 526]
- Complex filters use cross-coupling between real and imaginary signal paths to realize transfer functions that do not have the frequency-domain conjugate symmetry of real filters. [p. 527]
- Real filters can be frequency-translated into complex transfer functions by substituting $(s - j\omega)$ for s. [p. 528]
- Beginning with an active-RC prototype filter, duplicating it, and cross-coupling the two signal paths with resistors results in a complex frequency translation of the filter response. [p. 529]

12.12 REFERENCES

R. Allen. *Complex Analog Filters Obtained from Shifted Lowpass Prototypes*, M.A.Sc. Thesis, University of Toronto, 1985.

M. Banu and Y. Tsividis. "Fully Integrated Active RC Filters in MOS Technology," *IEEE J. of Solid-S tate Circuits,* Vol. 18, pp. 644–651, December 1983.

F. Behbahani, Y. Kishigami, J. Leete, and A. Abidi. "CMOS mixers and polyphase filters for large image rejection," *IEEE J. Solid-State Circuits*, vol. 36, pp. 873–887, June 2001.

K. Bult and H. Wallinga. "A Class of Analog CMOS Circuits Based on the Square-Law Characteristics of a MOS Transistor in Saturation," *IEEE J. of Solid-State Circuits,* Vol. 22, pp. 357–365, June 1987.

D. W. H. Calder. "Audio Frequency Gyrator Filters for an Integrated Radio Paging Receiver," in *Proc. 1984 IE E Conf. Mobile Radio Syst. Tech.,* pp. 21–26.

R. S. Carson. *Radio Communications Concepts: Analog.* John Wiley & Sons, New York, 1990.

Z. Chang, D. Macq, D. Haspeslagh, P. Spruyt, and B. Goffart. "A CMOS Analog Front-End Circuit for an FDM-Based ADSL System," *IEEE Int. Solid-State Circuits Conf.,* pp. 330–331, February 1995.

Z. Czarnul. "Modification of the Banu-Tsividis Continuous-Time Integrator Structure," *IEEE Trans. on Circuits and Systems,* Vol. 33, pp. 714–716, July 1986.

G. A. De Veirman and R. G. Yamasaki. "Design of a Bipolar 10-MHz Programmable Continuous-Time 0.05° Equiripple Linear Phase Filter," *IEEE J. of Solid-State Circuits,* Vol. 27, pp. 324–331, March 1992.

S. T. Dupuie and M. Ismail. "High-Frequency CMOS Transconductors," in *Analog IC Design: The Current-Mode Approach,* ed. C. Toumazou, F. J. Lidgey, and D. G. Haigh. Peter Peregrinus, London, UK, 1990.

A. Durham and W. Redman-White. "Integrated Continuous-Time Balanced Filters for 16-b DSP Interfaces," *IEEE J. of Solid-State Circuits,* Vol. 28, no. 7, pp. 835–839, July 1993.

B. Gilbert. "Current-Mode Circuits from a Translinear Viewpoint: A Tutorial," in *Analog IC Desig n: The Cur rent-Mode Approach,* ed. by C. Toumazou, F. J. Lidgey, and D. G. Haigh. Peter Peregrinus, London, UK, 1990.

M. J. Gingell, "Single-Sideband modulation using sequence asymmetric polyphase networks," *Elect. Com mun.*, vol. 48, no. 1–2, pp. 21–25, 1973.

H. Khorramabadi and P. R. Gray. "High-Frequency CMOS Continuous-Time Filters," *IEEE J. of Solid-State Circuits,* Vol. SC-19, no. 6, pp. 939–948, December 1984.

J. M. Khoury. "Design of a 15-MHz CMOS Continuous-Time Filter with On-Chip Tuning," *IEEE J. of Solid-State Circuits,* Vol. 26, pp. 1988–1997, December 1991.

M. Koyama, T. Arai, H. Tanimoto, and Y. Yoshida. "A 2.5-V Active Low-Pass Filter Using All npn Gilbert Cells with a 1-$V_{p\text{-}p}$ Linear Input Range," *IEEE J. of Solid-State Circuits,* Vol. 28, pp. 1246–1253, December 1993.

K. Kozma, D. A. Johns, and A. S. Sedra. "Automatic Tuning of Continuous-Time Integrated Filters Using an Adaptive Filter Technique," *IEEE Trans. on Circuits and Systems,* Vol. 38, pp. 1241–1248, November 1991.

F. Krummenacher and N. Joehl. "A 4-MHz CMOS Continuous-Time Filter with On-Chip Automatic Tuning," *IEEE J. of Solid-State Circuits,* Vol. 23, pp. 750–758, June 1988.

T. Kwan and K. Martin. "An Adaptive Analog Continuous-Time CMOS Biquadratic Filter," *IEEE J. of So lid-State Circuits,* Vol. 26, pp. 859–867, June 1991.

C. A. Laber and P. R. Gray. "A 20-MHz Sixth-Order BiCMOS Parasitic-Insensitive Continuous-Time Filter and Second-Order Equalizer Optimized for Disk-Drive Read Channels," *IEEE J. of Solid-State Circuits,* Vol. 28, pp. 462–470, April 1993.

K. R. Laker and W. M. C. Sansen. *Design of Analog Integrated Circuits and Systems.* McGraw-Hill, New York, 1994.

G. R. Lang and P.O. Brackett, "Complex Analogue Filters," *Proc. of the European Co nference o n Circuit Th eory a nd Design*, pp. 412–419, 1981.

K. Martin, "Complex signal processing is not complex" *IEEE Trans. Circuits and Systems I: Regu lar P apers*, vol. 51, pp. 1823–1836, Sept. 2004.

U. Moon and B.-S. Song. "Design of a Low-Distortion 22-kHz Fifth-Order Bessel Filter," *IEEE J. of Solid-S tate Circuits,* Vol. 28, pp. 1254–1264, December 1993.

J. P. Moree, G. Groenewold, and L. A. D. van den Broeke. "A Bipolar Integrated Continuous-Time Filter with Optimized Dynamic Range," *IEEE J. of Solid-State Circuits,* Vol. 28, pp. 954–961, September 1993.

A. Nedungadi and T. R. Viswanathan. "Design of Linear CMOS Transconductance Elements," *IEEE Trans. on Circuits and Systems,* Vol. 31, pp. 891–894, October 1984.

M. Padmanabhan and K. Martin. "A CMOS Analog Multi-Sinusoidal Phase-Locked-Loop," *IEEE J. of Solid-State Circuits,* Vol. 29, no. 9, pp. 1046–1057, September 1994.

C. S. Park and R. Schaumann. "A High-Frequency CMOS Linear Transconductance Element," *IEEE Trans. on Cir cuits and Systems,* Vol. 33, pp. 1132–1138, November 1986.

J. L. Pennock. "CMOS Triode Transconductor for Continuous-Time Active Integrated Filters," *Electron. Lett .,* Vol. 21, no. 18, pp. 817–818, August 1985.

R. Schaumann. "Continuous-Time Integrated Filters—A Tutorial," in *Integrated Continuous-Time Filters,* IEEE Press, 1992.

R. Schaumann, M. S. Ghausi, and K. R. Laker. *Design of Analog Filters,* Prentice Hall, Englewood Cliffs, New Jersey, 1990.

A. Sedra, W. Snelgrove, and R. Allen, "Complex analog bandpass filters designed by linearly shifting real low-pass prototypes", Proc. Int. Symp. Circuits and Systems, vol. 3, pp.1223–1226 1985.

E. Seevinck and R. F. Wassenaar. "A Versatile CMOS Linear Transconductor/Square-Law Function Circuit," *IEEE J. of Solid-State Circuits,* Vol. 22, pp. 366–377, June 1987.

A. Shoval, D. A. Johns, and W. M. Snelgrove. "A Wide-Range Tunable BiCMOS Transconductor," *Microelectronics J ournal,* Vol. 24, pp. 555–564, Elsevier Science Publishers, UK, August 1993.

A. Shoval, W. M. Snelgrove, and D. A. Johns. "A 100-Mb/s BiCMOS Adaptive Pulse-Shaping Filter," accepted for publication in *IEEE J. on Selected Areas in Communications: Special Issue on Copper Wire Access Technologies for High-Performance Networks,* 1995.

J. Silva-Martinez, M. S. J. Steyaert, and W. C. Sansen. "A Large-Signal Very Low-Distortion Transconductor for High-Frequency Continuous-Time Filters," *IEEE J. of Solid-State Circuits,* vol. 26, pp. 946–955, July 1991.

J. Silva-Martinez, M. S. J. Steyaert, and W. C. Sansen. "A 10.7-MHz 68-dB SNR CMOS Continuous-Time Filter with On-Chip Automatic Tuning," *IEEE J. of Solid-State Circuits,* Vol. 27, pp. 1843–1853, December 1992.

W.M. Snelgrove and A.S. Sedra. "State-Space Synthesis of Complex Analog Filters," *Proc. of the European Conference on Circuit Theory and Design*, pp. 420–424, 1981.

W. M. Snelgrove. *Intermediate Function Synthesis*, Ph.D. Thesis, University of Toronto, 1982.

W. M. Snelgrove and A. Shoval. "A Balanced 0.9-μm CMOS Transconductance-C Filter Tunable over the VHF Range," *IEEE J. of Solid-State Circuits,* Vol. 27, pp. 314–323, March 1992.

K. S. Tan and P. R. Gray. "Fully-Integrated Analog Filters Using Bipolar-JFET Technology," *IEEE J. of Solid-State Circuits,* Vol. SC-13, pp. 814–821, December 1978.

H. Tanimoto, M. Koyama, and Y. Yoshida. "Realization of a 1-V Active Filter Using a Linearization Technique Employing Plurality of Emitter-Coupled Pairs," *IEEE J. of Solid-State Circuits,* Vol. 26, pp. 937–945, July 1991.

Y. P. Tsividis. "Integrated Continuous-Time Filter Design—An Overview," *IEEE J. of Solid-State Circuits,* Vol. 29, pp. 166–176, March 1994.

Y. P. Tsividis and J. O. Voorman. *Integrated Continuous-Time Filter,* IEEE Press, New York, 1992.

T. R. Viswanathan, et al. "Switched-Capacitor Frequency Control Loop," *IEEE J. of Solid-State Circuits,* Vol. 17, pp. 775–778, August 1982.

Z. Wang and W. Guggenbühl. "A Voltage-Controllable Linear MOS Transconductor Using Bias Offset Technique," *IEEE J. of Solid-State Circuits,* Vol. 25, pp. 315–317, February 1990.

D. R. Welland, et al. "A Digital Read/Write Channel with EEPR4 Detection," *IEEE Int. Solid-State Circuits Conf.,* pp. 276–277, San Francisco, 1994.

S. D. Willingham, K. W. Martin, and A. Ganesan. "A BiCMOS Low-Distortion 8-MHz Low-Pass Filter," *IEEE J. of Solid-State Circuits,* Vol. 28, pp. 1234–1245, December 1993.

S. L. Wong. "Novel Drain-Based Transconductance Building Blocks for Continuous-Time Filter Applications," *Electron. Lett.,* Vol. 25, no. 2, pp. 100–101, January 1989.

A. Wyszynski, R. Schaumann, S. Szczepanski, and P. Van Halen. "Design of a 2.7-GHz Linear OTA and a 250-MHz Elliptic Filter in Bipolar Transistor-Array Technology," *IEEE Trans. on Circuits and Systems—II: Analog and Digital Signal Processing,* Vol. 40, pp. 19–31, January 1993.

12.13 PROBLEMS

Unless otherwise stated, assume the 0.8 μm technology CMOS devices from Table 1.5, and npn bipolar transistors with the following parameters:

$\beta = 100$
$V_A = 80$ V
$\tau_b = 13$ ps
$\tau_s = 4$ ns
$r_b = 330\ \Omega$

12.13.1 Section 12.1: Introduction to Continuous-Time Filters

12.1 Find all of the signal path gains that are required in the general biquad in Fig. 12.2 to realize a bandpass filter with a center frequency of $\omega_0 = 2\pi \cdot 34$ MHz and a Q-factor of 3.

12.2 Find the transfer function of the general biquadratic signal flow graph in Fig. P12.2.

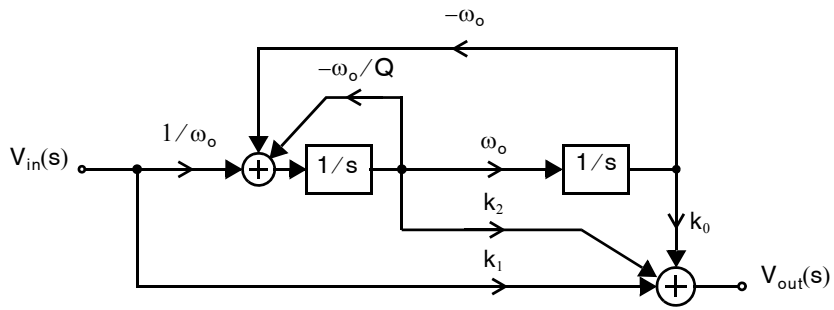


Fig. P12.2

12.13.2 Section 12.2: Introduction to G_m-C Filters

12.3 Draw the schematic of a fully-differential G_m-C implementation of the signal flow graph in Fig. P12.2.

12.4 Based on the block diagram of Fig. 12.9, find transconductance values for a first-order filter with a dc gain of 10 and a pole at 15 MHz (no finite zero). Assume the integrating capacitors are sized $C_A = 5$ pF.

12.5 Find the transconductance and capacitance values of the second-order filter of Fig. 12.10 for a low-pass filter with a pole frequency of 10 MHz, $Q = 1$, and a dc gain of 5.

12.6 Derive the design equations for the biquad circuit shown in Fig. 12.18.

12.13.3 Sections 12.3 to 12.7: Transconductor Circuits

12.7 Consider the transconductor shown in Fig. 12.13. Find the transconductance of this circuit for finite opamp gains A.

12.8 In the transconductor of Fig. 12.12(*b*), find the minimum collector voltages of Q_1, Q_2 such that they remain in the active region for $v_i = \pm 1$ V around a common-mode voltage of 2 V. Repeat for the transconductor of Fig. 12.13 where Q_1 and Q_2 are replaced by bipolar transistors.

12.9 In the transconductor of Fig. 12.14, find the new transconductance value if the base-emitter areas of Q_7, Q_8 are four times those of Q_3, Q_4.

12.10 Consider the transconductor of Fig. 12.14, where we want $G_m = 1$ mA/V. What are reasonable values for I_1 and I_2 such that all transistors do not fall below 20 percent of their nominal bias current when the peak value of $v_i = v_i^+ - v_i^-$ is 500 mV?

12.11 Consider the linearized two-differential pairs shown in Fig. 12.36, where $I_1 = 2$ mA. Find the percentage of error between an ideal linear transconductor and the true output current when $v_i = 48$ mV.

12.12 Consider the circuit Fig. 12.15, with $I_1 = 100\ \mu$A, $I_2 = 10\ \mu$A, $(W/L)_9 = 2$, $V_C = 5$ V, and $v_i^- = 2.5$ V.

- **a.** Find the transconductance, G_m.
- **b.** Assuming a perfectly linear transconductor, find i_{o1} when v_i^+ equals 2.6 V and 3 V.
- **c.** Find the true value of i_{o1} when v_i^+ equals 2.6 V and 3 V by taking into account the triode equation for Q_9 (assume everything else is ideal). Compare your results with those in part (b).

12.13 Choose reasonable transistor sizes for the circuit in Fig. 12.19 to realize a transconductor that has $G_m = 0.3$ mA/V, $k_1/k_3 = 6.7$, and a peak differential-input voltage of 1 V.

12.14 Consider the circuit in Fig. 12.24, where all transistors have $W/L = 10\ \mu$m$/2\ \mu$m, $V_{DD} = -V_{SS} = 2.5$ V, and $V_{C1} = -V_{C2} = 2.0$ V. Find the transconductance of the circuit.

12.15 Show that the transconductance of Fig. 12.26 is given by

$$G_m = \left(\frac{n}{n+1}\right) 4\sqrt{kI_B}$$

12.16 Consider the transconductor shown in Fig. 12.28, where all transistors have size $W/L = 10\ \mu m/2\ \mu m$, $I_B = 50\ \mu A$, and $V_{DD} = -V_{SS} = 3\ V$. What is the maximum differential input, $v_1 - v_2$ (centered around 0 V), that can be used while maintaining proper operation?

12.17 Show that the circuit of Fig. 12.30 results in

$$i_1 - i_2 = K\left(V_B - 2V_{tn}\right)\left(2v_1 - V_B\right)$$

12.13.4 Section 12.8: Active RC and MOSFET-C Filters

12.18 Draw a signal flow graph representation of the active RC filter in Fig. P12.18 with signal path gains written in terms of the passive component values.

12.19 Find the frequency response of the active RC filter in Fig. P12.18 in terms of the resistor and capacitor component values assuming an ideal opamp.

12.20 Repeat Problem 12.19, this time assuming the opamp has a finite dc gain A_0 and unity-gain frequency ω_{ta}. Sketch the resulting magnitude response $|V_{in}(\omega)/V_{out}(\omega)|$ and compare it to that expected with an ideal opamp.

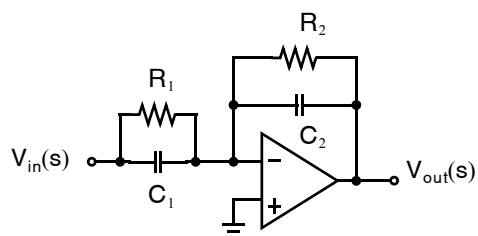


Fig. P12.18

12.21 Design a first order lowpass active RC filter based upon the circuit of Fig. P12.18 to have a dc gain of 3 and a 3-dB bandwidth of 15 MHz.

12.22 Draw the schematic of a fully-differential two-transistor MOSFET-C filter implementation of the signal flow graph in Fig. P12.2.

12.23 Using two-transistor integrators, design a MOSFET-C second-order low-pass filter with a pole frequency of $1\ MHz$, $Q = 1$, and a dc gain of 2. Let the integrating capacitors be $10\ pF$, $V_C = 3\ V$, and $V_x = 0\ V$.

CHAPTER

13 Discrete-Time Signals

A basic understanding of discrete-time signal processing is essential in the design of most modern analog systems. For example, discrete-time signal processing is heavily used in the design and analysis of oversampling analog-to-digital (A/D) and digital-to-analog (D/A) converters used in digital audio, wireless communication, and instrumentation applications. Also, a discrete-time filtering technique known as switched-capacitor filtering is a popular approach for realizing fully integrated analog filters. Switched-capacitor filters are in the class of analog filters since voltage levels in these filters remain continuous. In other words, switched-capacitor filters operate and are analyzed using discrete time-steps but involve no A/D or D/A converters; hence, they are quantized in time but not in amplitude. This chapter presents some basic concepts of discrete-time signals and filters.

13.1 OVERVIEW OF SOME SIGNAL SPECTRA

Consider the spectra of sampled and continuous-time signals in the block diagram systems shown in Fig. 13.1, where it is assumed that the continuous-time signal, $x_c(t)$, is band limited through the use of an anti-aliasing filter (not shown). *DSP* refers to *Discrete-time signal processing*, which may be accomplished using fully digital processing or discrete-time analog circuits such as switched-capacitor filters. Some example time signals and frequency spectra for this system are shown in Fig. 13.2. Here, $s(t)$ is a periodic impulse train in time with a period of T, where T equals the inverse of the sampling frequency, f_s. Some relationships for the signals in Figs. 13.1 and 13.2 are as follows:

Key Point: *Discrete-time signals may be processed either digitally, in which case they are quantized both in time and amplitude, or using discrete-time analog circuits such as switched-capacitor circuits, in which case the signal amplitude is continuously variable.*

1. $x_s(t)$ has the same frequency spectrum as $x_c(t)$, but the baseband spectrum repeats every f_s (assuming that no aliasing occurs and hence, an anti-aliasing filter is needed).
2. $x(n)$ has the same frequency spectrum as $x_s(t)$, but the sampling frequency is normalized to 1.
3. The frequency spectrum for $x_{sh}(t)$ equals that of $x_s(t)$ multiplied by a response of $(\sin x)/x$ (in effect, multiplying $X_s(f)$ by this response helps to remove high-frequency images).

The remainder of this chapter confirms these spectral relationships and introduces other basic discrete-time concepts.

13.2 LAPLACE TRANSFORMS OF DISCRETE-TIME SIGNALS

Consider the sampled signal, $x_s(t)$, related to the continuous-time signal, $x_c(t)$, as shown in Fig. 13.3. The signal $x_s(t)$ is fictitious; it does not arise anywhere in a practical physical realization, such as the one depicted in Fig. 13.1(*b*). However, it is a useful tool to aid our understanding and modeling of practical discrete-time signals.

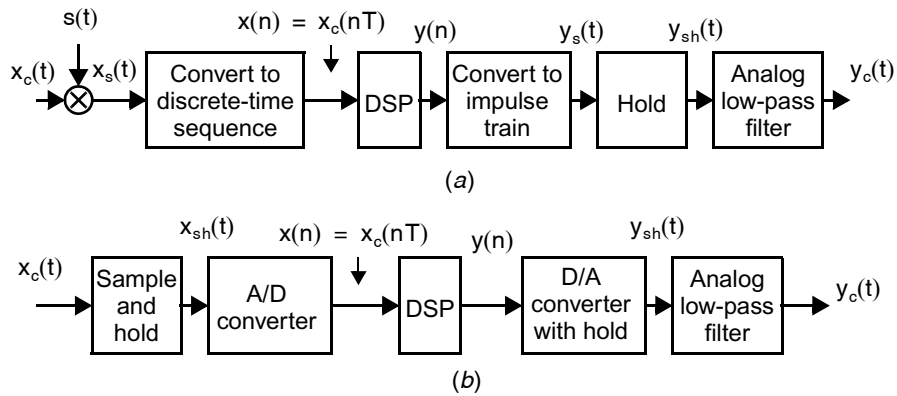


Fig. 13.1 Performing DSP on ana log signals. (*a*) C onceptual realization, a nd (*b*) typ ical physical realization.

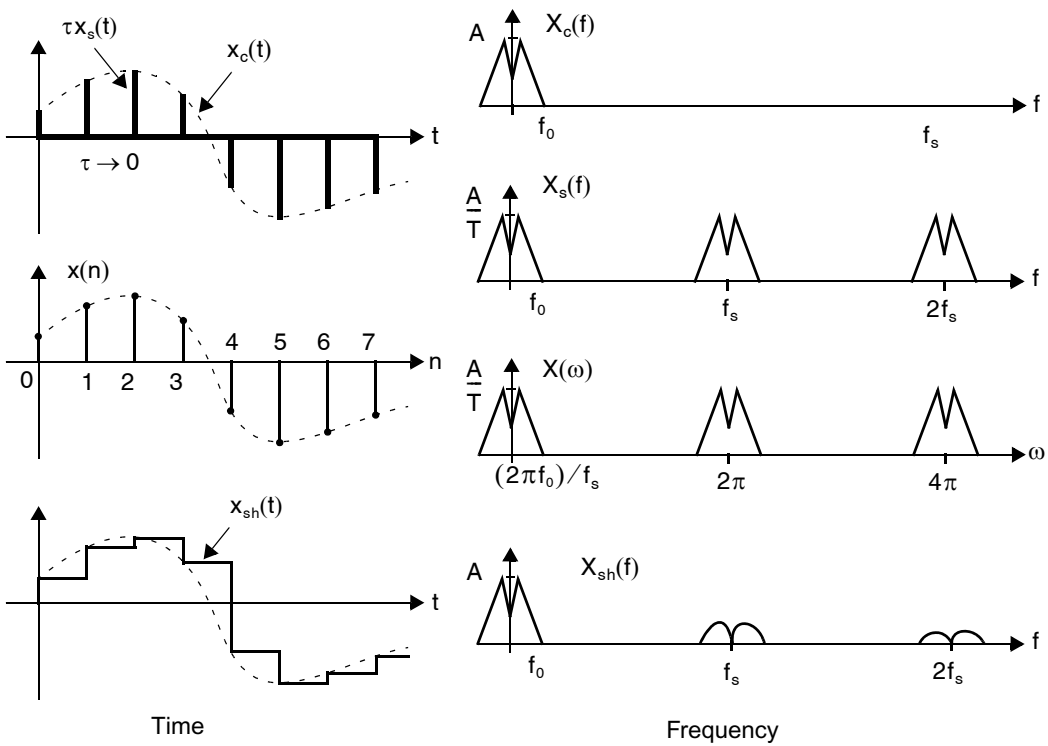


Fig. 13.2 Some time signals and frequency spectra.

Compared to $x_c(t)$, $x_s(t)$ has been scaled by τ such that the area under the pulse at nT equals the value of $x_c(nT)$. In other words, at $t = nT$, we have

$$x_s(nT) = \frac{x_c(nT)}{\tau} \tag{13.1}$$

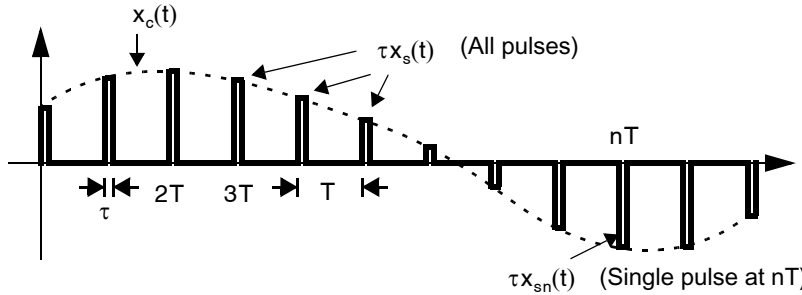


Fig. 13.3 Sampled and continuous-time signals.

such that the area under the pulse, $\tau x_s(nT)$, equals $x_c(nT)$. Thus, as $\tau \to 0$, the height of $x_s(t)$ at time nT goes to ∞, and so we plot $\tau x_s(t)$ instead of $x_s(t)$.

We define $\vartheta(t)$ to be the step function given by

$$\vartheta(t) \equiv \begin{cases} 1 & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{13.2}$$

Then $x_s(t)$ can be represented as a linear combination of a series of pulses, $x_{sn}(t)$, where $x_{sn}(t)$ is zero everywhere except for a single pulse at nT. The single-pulse signal, $x_{sn}(t)$, can be written as

$$x_{sn}(t) = \frac{x_c(nT)}{\tau}[\vartheta(t - nT) - \vartheta(t - nT - \tau)] \tag{13.3}$$

so that we can now write $x_s(t)$ as

$$x_s(t) = \sum_{n=-\infty}^{\infty} x_{sn}(t) \tag{13.4}$$

Note that these signals are defined for *all time,* so we can find the Laplace transform of $x_s(t)$ in terms of $x_c(t)$. Using the notation that $X_s(s)$ is the Laplace transform of $x_s(t)$, we find the Laplace transform $X_{sn}(s)$ for $x_{sn}(t)$ to be given by

$$X_{sn}(s) = \frac{1}{\tau}\left(\frac{1 - e^{-s\tau}}{s}\right) x_c(nT)e^{-snT} \tag{13.5}$$

Since $x_s(t)$ is merely a linear combination of $x_{sn}(t)$, we also have (for $\tau \neq 0$)

$$X_s(s) = \frac{1}{\tau}\left(\frac{1 - e^{-s\tau}}{s}\right) \sum_{n=-\infty}^{\infty} x_c(nT)e^{-snT} \tag{13.6}$$

Using the expansion $e^x = 1 + x + (x^2/2!) + \cdots$, when $\tau \to 0$, the term before the summation in (13.6) goes to unity. Therefore, in the limiting case as $\tau \to 0$, we have

$$X_s(s) = \sum_{n=-\infty}^{\infty} x_c(nT)e^{-snT} \tag{13.7}$$

13.2.1 Spectra of Discrete-Time Signals

The spectrum of the sampled signal, $x_s(t)$, can be found by replacing s by $j\omega$ in (13.7). However, a more intuitive approach to find the spectrum of $x_s(t)$ is to recall that multiplication in the time domain is equivalent to convolution in the frequency domain. To use this fact, note that, for $\tau \to 0$, $x_s(t)$ can be written as the product

$$x_s(t) = x_c(t)s(t) \tag{13.8}$$

where $s(t)$ is a periodic pulse train, or mathematically,

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{13.9}$$

where $\delta(t)$ is the unit impulse function, also called the *Dirac delta function*. It is well known that the Fourier transform of a periodic impulse train is another periodic impulse train. Specifically, the spectrum of $s(t)$, $S(j\omega)$, is given by

$$S(j\omega) = \frac{2\pi}{T}\sum_{k=-\infty}^{\infty} \delta\left(\omega - k\frac{2\pi}{T}\right) \tag{13.10}$$

Now writing (13.8) in the frequency domain, we have

$$X_s(j\omega) = \frac{1}{2\pi}X_c(j\omega) \otimes S(j\omega) \tag{13.11}$$

where $\otimes$ denotes convolution. Finally, by performing this convolution either mathematically or graphically, the spectrum of $X_s(j\omega)$ is seen to be given by

$$X_s(j\omega) = \frac{1}{T}\sum_{k=-\infty}^{\infty} X_c\left(j\omega - \frac{jk2\pi}{T}\right) \tag{13.12}$$

or, equivalently,

$$X_s(f) = \frac{1}{T}\sum_{k=-\infty}^{\infty} X_c(j2\pi f - jk2\pi f_s) \tag{13.13}$$

Key Point: *The spectrum of a sampled signal is a sum of shifted copies of the original continuous-time spectrum so that no aliasing occurs if the original waveform is bandlimited to one-half the sampling rate. The minimum sampling frequency required to avoid aliasing is called the Nyquist rate, and is equal to two times the signal bandwidth.*

Equations (13.12) and (13.13) show that the spectrum of the sampled signal, $x_s(t)$, equals a sum of shifted spectra of $x_c(t)$, and therefore no aliasing occurs if $X_c(j\omega)$ is band limited to $f_s/2$. The minimum sampling frequency required to avoid aliasing is called the Nyquist rate, and is equal to two times the signal bandwidth. The relation in (13.13) also confirms the example spectrum for $X_s(f)$, shown in Fig. 13.2. Each copy of the dirac-delta-sampled spectrum is identical: $X_s(f) = X_s(f \pm kf_s)$, where k is an arbitrary integer as seen by substitution in (13.13).[1]

Finally, note that the signal $x_s(t)$ cannot exist in practice when $\tau \to 0$ since an infinite amount of power would be required to create it. This is evident when one considers the integral of $X_s(f)$ over all frequencies.

1. The factor $j2\pi$ is sometimes omitted from the argument of continuous-time spectra for convenience.

13.3 z-TRANSFORM

For our purposes, the z-transform is merely a shorthand notation for (13.7). Specifically, defining

$$z \equiv e^{sT} \tag{13.14}$$

we can write

$$X(z) \equiv \sum_{n=-\infty}^{\infty} x_c(nT)z^{-n} \tag{13.15}$$

where $X(z)$ is called the z-transform of the samples $x_c(nT)$.

Two properties of the z-transform that can be deduced from Laplace transform properties are as follows:

1. If $x(n) \leftrightarrow X(z)$, then $x(n-k) \leftrightarrow z^{-k}X(z)$
2. Convolution in the time domain is equivalent to multiplication in the frequency domain. Specifically, if $y(n) = h(n) \otimes x(n)$, where $\otimes$ denotes convolution, then $Y(z) = H(z)X(z)$. Similarly, multiplication in the time domain is equivalent to convolution in the frequency domain.

Note that $X(z)$ is not a function of the sampling rate but is related only to the numbers $x_c(nT)$, whereas $X_s(s)$ is the Laplace transform of the signal $x_s(t)$ as $\tau \to 0$. In other words, the signal $x(n)$ is simply a series of numbers that may (or may not) have been obtained by sampling a continuous-time signal. One way of thinking about this series of numbers as they relate to the samples of a possible continuous-time signal is that the original sample time, T, has been effectively normalized to one (i.e., $f_s' = 1$ Hz). Such a normalization of the sample time, T, in both time and frequency, justifies the spectral relation between $X_s(f)$ and $X(\omega)$ shown in Fig. 13.2. Specifically, the relationship between $X_s(f)$ and $X(\omega)$ is given by

$$X_s(f) = X\left(\frac{2\pi f}{f_s}\right) \tag{13.16}$$

or, equivalently, the following frequency scaling has been applied:

$$\omega = \frac{2\pi f}{f_s} \tag{13.17}$$

This normalization results in discrete-time signals having ω in units of radians/sample, whereas the original continuous-time signals have frequency units of cycles/second (hertz) or radians/second. For example, a continuous-time sinusoidal signal of 1 kHz when sampled at 4 kHz will change by $\pi/2$ radians between each sample. Therefore, such a discrete-time signal is defined to have a frequency of $\pi/2$ rad/sample. Other examples of discrete-time sinusoidal signals are shown in Fig. 13.4. The spectra of discrete-time signals are always invariant under shifts of 2π radians/sample. For example, a discrete-time signal that has a frequency of $\pi/4$ rad/sample is identical to that of $9\pi/4$ rad/sample. Thus, normally discrete-time signals are completely specified by considering their spectra only between $-\pi$ and π rad/sample. For a more detailed discussion of this topic, see [Proakis, 1992].

> ***Key Point:*** *Discrete-time signal spectra are equal to the corresponding sampled signal spectra with the sample time normalized to 1 Hz. Hence, their spectra repeat every 2π.*

EXAMPLE 13.1

Consider the spectra of $X_c(f)$ and $X_s(f)$, shown in Fig. 13.2, where f_0 is 1 Hz and f_s is 4 Hz. Compare the time and spectrum plots of $X_s(f)$ and $X_{s2}(f)$, where $X_{s2}(f)$ is sampled at 12 Hz. How do the spectra differ between the two sampling rates?

Solution

By sampling at 12 Hz, the spectrum of $X_c(f)$ repeats every 12 Hz, resulting in the signals shown in Fig. 13.5.

Note that, for $X(\omega)$, 4 Hz is normalized to 2π rad/sample, whereas for $X_2(\omega)$, 12 Hz is normalized to 2π rad/sample.

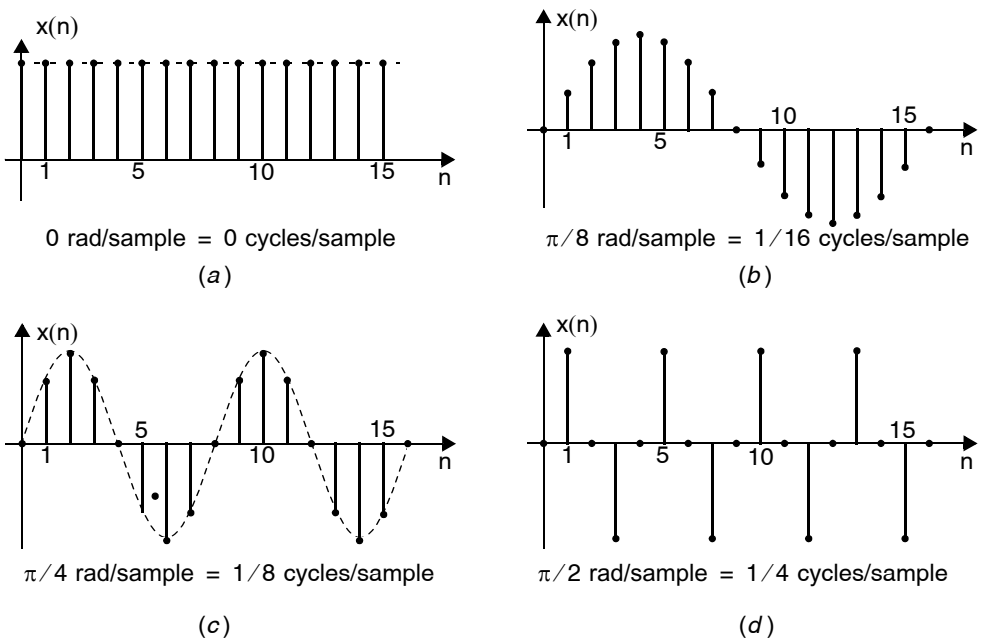


Fig. 13.4 Some discrete-time sinusoidal signals.

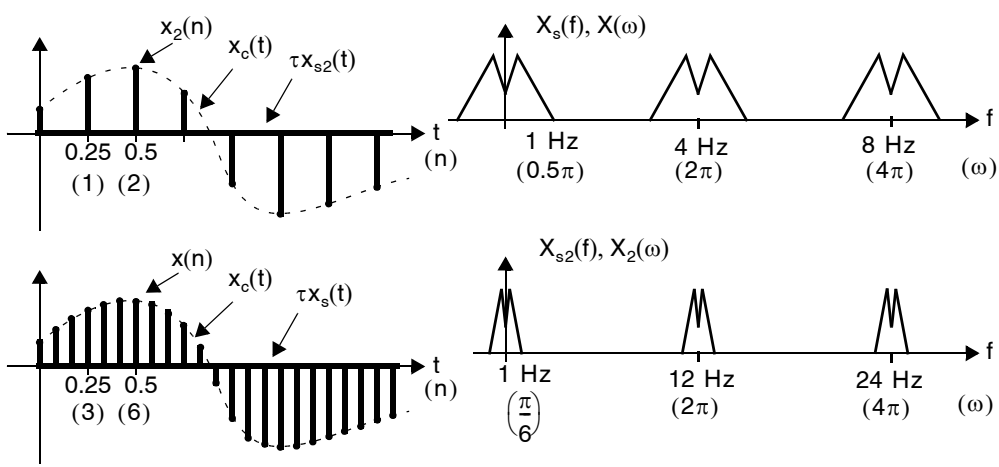


Fig. 13.5 Comparing time and frequency of two sampling rates.

13.4 DOWNSAMPLING AND UPSAMPLING

Two important operations in discrete-time signal processing are downsampling and upsampling. Downsampling is used to reduce the sample rate, and hence the amount of information that must be processed and/or stored, hopefully without information loss. Upsampling is used to increase the sample rate. Although noninteger downsampling and upsampling rates can be achieved, here we consider only the case in which L is an integer value.

Downsampling is achieved by keeping every Lth sample and discarding the others. As Fig. 13.6 shows, the result of downsampling is to expand the original spectra by L. Thus, to avoid digital aliasing, the spectrum of the original signal must be band limited to π/L before downsampling is done. In other words, the signal must be sampled L times above its minimum sampling rate so that no information is lost during downsampling.

Key Point: *To avoid aliasing when downsampling by a factor L, the original signal must be sampled at L times the Nyquist rate.*

Upsampling is accomplished by inserting $L - 1$ zero values between each pair of consecutive samples, as shown in Fig. 13.7. In this case, one can show that the spectrum of the resulting upsampled signal is identical to that of the original signal but with a renormalization along the frequency axis. Specifically, when a signal is upsampled by L, the frequency axis is scaled by L such that 2π now occurs where $L2\pi$ occurred in the original signal and L identical copies of the original spectrum are now squeezed within the range 0 to 2π. This operation is useful when one wishes to increase the effective sampling rate of a signal, particularly if postfiltering is then applied.

Key Point: *Upsampling by L results in a renormalization along the frequency axis so that L copies of the original signal spectrum are now squeezed in the range 0 to 2π.*

Follow the next two examples carefully—they help explain the spectral changes due to downsampling and upsampling.

EXAMPLE 13.2

The signal $x_2(n)$ in Example 13.1 is to be downsampled by $L = 3$. Find the new spectrum, $X_3(\omega)$, when every third sample is kept and the others are discarded. In other words, if the original series of numbers is $x_1, x_2, x_3, x_4, \ldots$, the new series is $x_1, x_4, x_7, x_{10}, \ldots$.

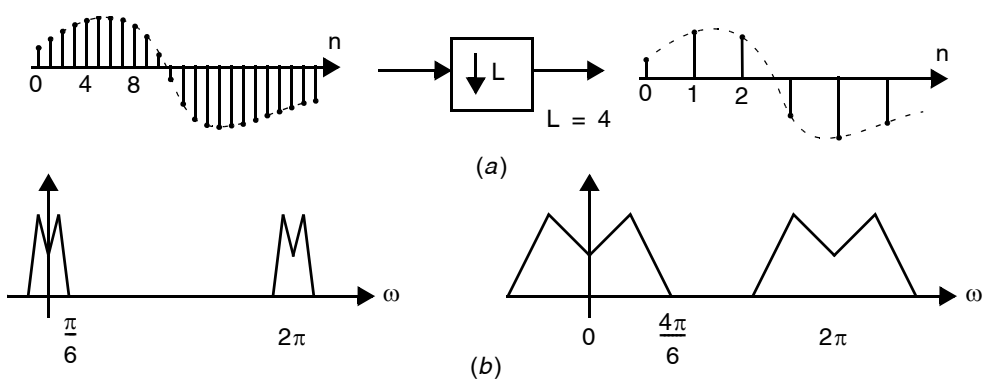


Fig. 13.6 Downsampling by 4: (*a*) time domain, and (*b*) frequency domain.

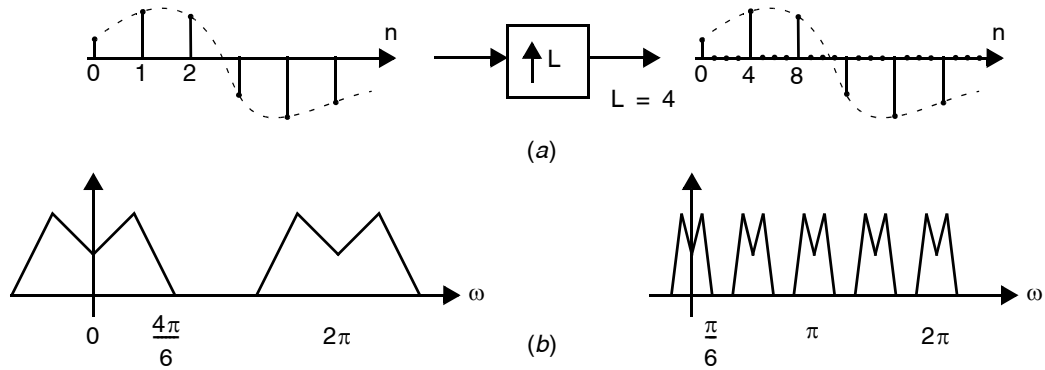


Fig. 13.7 Upsampling by 4: (*a*) time domain, and (*b*) frequency domain.

Solution

Clearly the solution here is that the new sequence is equivalent to $x(n)$ in Example 13.1, and thus its spectrum is simply $X_3(\omega) = X(\omega)$.

Note that no information was lost in this example because the initial images of the signal $x_2(n)$ were far enough apart. However, if the downsampling operation results in the spectra overlapping one another, then an aliasing phenomenon occurs that is similar to that which occurs when sampling an analog signal lower than the Nyquist rate. In fact, it can be shown that, to avoid any aliasing during downsampling by the factor L, the original signal should be band limited to π/L.

EXAMPLE 13.3

As an example of upsampling of the signal $x(n)$ in Example 13.1, find the new spectrum, $X_4(\omega)$, if two zeros are inserted between each sample (i.e., if we upsample by $L = 3$). In other words, if the original series of numbers was $x(n) = x_1, x_2, x_3, \ldots$ and had a spectrum $X(\omega)$, then the new series of numbers is now $x_4(n) = (x_1, 0, 0, x_2, 0, 0, x_3, \ldots)$ and has a spectrum $X_4(\omega)$.

Solution

Perhaps the simplest way to conceptualize the insertion of these extra 0 samples in a series is to look at $X_{s4}(f)$ in time and frequency when we let $f_{s4} = 12$ Hz. Recall that $x_{s4}(t)$ is defined for all time, where it is zero between its impulses, and the Laplace transform is used to observe its frequency response. With such a sampling frequency, we note that $x_{s4}(t)$ is equal to the signal $x_s(t)$, and thus $X_{s4}(f) = X_s(f)$. To find $X_4(\omega)$, recall that it is simply a frequency normalization of $X_{s4}(f)$, where the sampling frequency is normalized to 2π. In other words, the series $x(n)$ is simply a normalization of the time between impulses to 1. However, by inserting extra zeros between samples, the normalization is effectively being done for a time period smaller than that between nonzero impulses. Thus, the images remain the same while the normalization along the frequency axis is different.

In this example, since two zeros are inserted, the effective sampling rate might be thought of as 12 Hz, but now the images at 4 Hz and 8 Hz remain in contrast to $X_2(\omega)$ in Example 13.1. The resulting signals are shown in Fig. 13.8.

Finally, note that if signal processing is used on the samples $x_4(n)$ to eliminate the two images at $(2\pi)/3$ and $(4\pi)/3$, then the resulting signal will equal $x_2(n)$. In other words, as long as an analog signal is originally sampled higher than the Nyquist rate, we can use upsampling and digital-signal processing to derive signals that are the same as if the analog signal were sampled at a much higher rate.

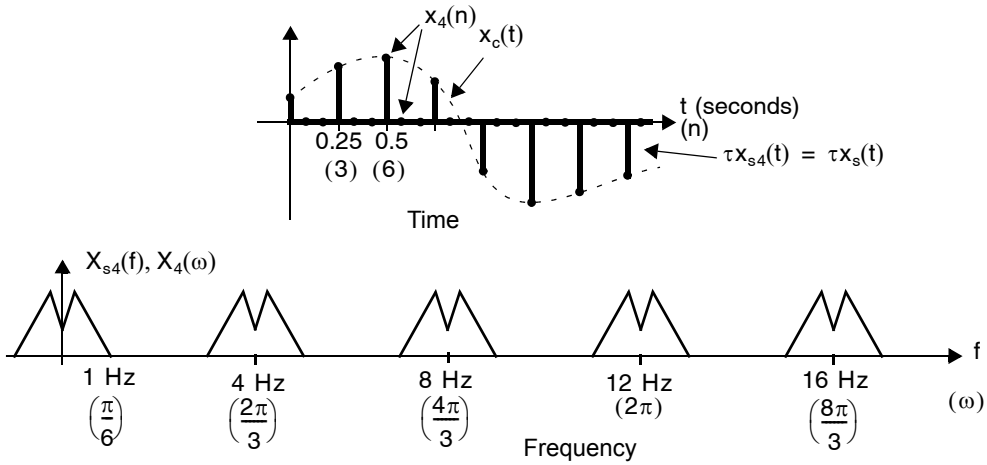


Fig. 13.8 Spectrum of a discrete-time signal upsampled by 3.

13.5 DISCRETE-TIME FILTERS

Thus far, we have seen the relationship between continuous-time and discrete-time signals in the time and frequency domain. However, often one wishes to perform filtering on a discrete-time signal to produce another discrete-time signal. In other words, an input series of numbers is applied to a discrete-time filter to create an output series of numbers. This filtering of discrete-time signals is most easily visualized with the shorthand notation of z-transforms.

Consider the system shown in Fig. 13.9, where the output signal is defined to be the impulse response, h(n), when the input, u(n), is an impulse (i.e., 1 for n = 0 and 0 otherwise). The transfer function of the filter is said to be given by H(z), which is the z-transform of the impulse response, h(n).

13.5.1 Frequency Response of Discrete-Time Filters

The transfer functions for discrete-time filters appear similar to those for continuous-time filters; practical integrated circuit discrete-time transfer functions are represented by rational transfer functions except that, instead of polynomials in s, polynomials in z are obtained. For example, the transfer function of a low-pass, continuous-time filter, $H_c(s)$, might appear as

$$H_c(s) = \frac{4}{s^2 + 2s + 4} \tag{13.18}$$

The poles for this filter are determined by finding the roots of the denominator polynomial, which are $-1.0 \pm 1.7321j$ for this example. This continuous-time filter is also defined to have two zeros at ∞ since the

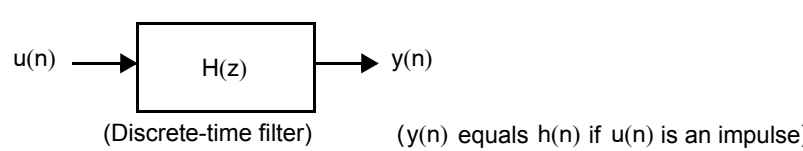


Fig. 13.9 Discrete-time filter system.

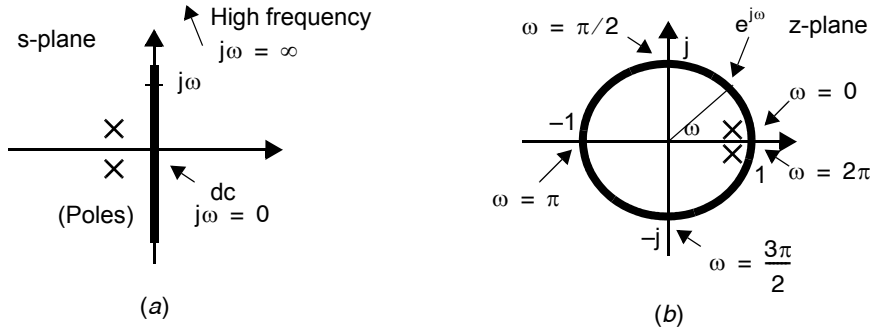


Fig. 13.10 Transfer function response. (*a*) Continuous-time, and (*b*) discrete-time.

denominator polynomial is two orders higher than the numerator polynomial. To find the frequency response of $H_c(s)$, it is evaluated along the $j\omega$ axis by substituting $s = j\omega$. The poles and zeros are plotted on the s-plane in Fig. 13.10(*a*), and the magnitude and phase response $H_c(j\omega)$ can be found in terms of the magnitude and phase of vectors from all poles (and finite zeros) to the point $s = j\omega$ along the frequency-axis.

$$|H_c(j\omega)| = \frac{4}{|j\omega + 1.0 + 1.7321j||j\omega + 1.0 - 1.7321j|} \tag{13.19}$$

$$\angle H_c(j\omega) = -\angle(j\omega + 1.0 + 1.7321j) - \angle(j\omega + 1.0 - 1.7321j) \tag{13.20}$$

An example of a discrete-time, low-pass transfer function is given by the following equation:

$$H(z) = \frac{0.05}{z^2 - 1.6z + 0.65} \tag{13.21}$$

Key Point: *The frequency response of a discrete-time filter H(z) is found by evaluating its magnitude and phase on the unit circle, or graphically by considering the magnitude and phase of vectors connecting its poles and zeros to points around the unit circle.*

The poles of $H(z)$ occur at $0.8 \pm 0.1j$ in the z-plane, and two zeros are again at ∞. To find the frequency response of $H(z)$, we evaluate it around the unit circle contour $z = e^{j\omega}$ as shown in Fig. 13.10(*b*), instead of going along the vertical $j\omega$ axis as in the s-plane. Note that substituting $z = e^{j\omega}$ in the z-domain transfer function, $H(z)$, is simply a result of substituting $s = j\omega$ into (13.14), where T has been normalized to 1, as discussed in Section 13.3. The discrete-time frequency response can be found graphically by considering the magnitude and phase of vectors from each pole and zero of $H(z)$ as shown in Fig. 13.10(*b*) to the corresponding frequency on the unit circle $z = e^{j\omega}$. Poles or zeros occurring at $z = 0$ do not affect the magnitude response $H(z)$ since a vector from the origin to the unit circle always has a length of unity. However, they do affect the phase response.

Key Point: *In discrete-time, the point $z = 1$ corresponds both to dc and to signals at the sampling frequency. The frequency response is periodic, and for rational transfer functions with real-valued coefficients exhibits conjugate symmetry around dc. Hence, discrete-time frequency responses need only be plotted from dc to half the sample rate, π.*

In discrete time, $H(z = 1)$ corresponds to the frequency response at both dc (i.e., $\omega = 0$) and at $\omega = 2\pi$. Also, the time normalization of setting T to unity implies that $\omega = 2\pi$ is equivalent to the sampling-rate speed (i.e., $f = f_s$) for $X_s(f)$. In addition, note that the frequency response of a filter need be plotted only for $0 \le \omega \le \pi$ (i.e., $0 \le \omega \le f_s/2$) since, for filters with real coefficients, the poles and zeros always occur in complex-conjugate pairs (or on the real axis), so the magnitude response of the filter is equal to that for $\pi \le \omega \le 2\pi$ (the filter's phase is antisymmetric). Going around the circle again gives the same result as the first time, implying that the frequency response repeats every 2π.

Before we leave this section, a word of caution is in order. To simplify notation, the same variables, f and ω, are used in both the continuous-time and discrete-time domains in Fig. 13.10. However, these variables are not equal in the two domains, and care should be taken not to confuse values from the two domains. The continuous-time domain is used here for illustrative reasons only since the reader should already be quite familiar with transfer function analysis in the s-domain. In summary, the unit circle, $e^{j\omega}$, is used to determine the frequency response of a system that has its input and output as a series of numbers, whereas the $j\omega$-axis is used for a system that has continuous-time inputs and outputs. However, ω is different for the two domains.

EXAMPLE 13.4

Assuming a sample rate of $f_s = 100$ kHz, find the magnitude of the transfer function in (13.21) for 0 Hz, 100 Hz, 1 kHz, 10 kHz, 50 kHz, 90 kHz, and 100 kHz.

Solution

To find the magnitude of $H(z)$ at these frequencies, first we find their equivalent z-domain locations by normalizing f_s to 2π. Next, we find the gain of $H(z)$ by putting these z-domain values into (13.21) and finding the magnitude of the resulting complex value. Table 13.1 summarizes the results.

Note that the gain of $H(z)$ is the same at both 0 Hz and 100 kHz, as expected, since their z-plane locations are the same. Also, the gain is the same at both 10 kHz and 90 kHz since their z-plane locations are complex conjugates of each other. Finally, note that the minimum gain for this transfer function occurs at $z = -1$, or equivalently, $f_s/2 = 50$ kHz.

EXAMPLE 13.5

Consider a first-order $H(z)$ having its zero at ∞ and its pole on the real axis, where $0 < a < 1$. Mathematically, the transfer function is represented by $H(z) = b/(z - a)$. Find the value of ω, where the magnitude of $H(z)$ is 3 dB lower than its dc value. What is the 3-dB value of ω for a real pole at 0.8? What fraction of the sampling rate, f_s, does it correspond to?

Table 13.1 Finding the gain of H(z) at some example frequencies.

	z-plane locations		
Frequency (kHz)	$e^{j\omega}$	$z = x + jy$	$\lvert H(z)\rvert$
0	e^{j0}	1.0 + j0.0	1.0
0.1	$e^{j0.002\pi}$	0.9999803 + j0.00628314	0.9997
1	$e^{j0.02\pi}$	0.9980267 + j0.06279052	0.968
10	$e^{j0.2\pi}$	0.809 + j0.5878	0.149
50	$e^{j\pi}$	−1.0 + j0.0	0.0154
90	$e^{j1.8\pi}$	0.809 − j0.5878	0.149
100	$e^{j2\pi}$	1.0 + j0.0	1.0

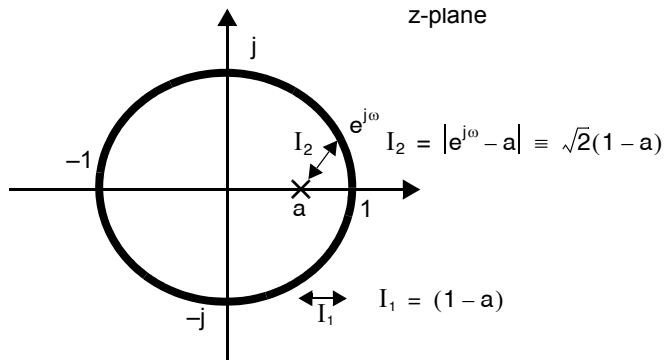


Fig. 13.11 Pole-zero plot used to determine the 3-dB frequency of a first-order, discrete-time filter.

Solution

Consider the pole-zero plot shown in Fig. 13.11, where the zero is not shown since it is at ∞. Since the zero is at ∞, we need be concerned only with the magnitude of the denominator of $H(z)$ and when it becomes $\sqrt{2}$ larger than its dc value. The magnitude of the denominator of $H(z)$ for $z = 1$ (i.e., dc) is shown as the vector I_1 in Fig. 13.11. This vector changes in size as z goes around the unit circle and is shown as I_2 when it becomes $\sqrt{2}$ larger than I_1. Thus we can write

$$I_2 = |e^{j\omega} - a| \equiv \sqrt{2}(1-a) \tag{13.22}$$

Writing $e^{j\omega} = \cos(\omega) + j\sin(\omega)$, we have

$$|(\cos(\omega) - a) + j\sin(\omega)|^2 = 2(1-a)^2 \tag{13.23}$$

$$\cos^2(\omega) - (2a)\cos(\omega) + a^2 + \sin^2(\omega) = 2(1-a)^2 \tag{13.24}$$

$$1 - (2a)\cos(\omega) + a^2 = 2(1 - 2a + a^2) \tag{13.25}$$

which is rearranged to give the final result.

$$\omega = \cos^{-1}\left(2 - \frac{a}{2} - \frac{1}{2a}\right) \tag{13.26}$$

For $a = 0.8$, $\omega = 0.2241$ rad, or 12.84 degrees. Such a location on the unit circle corresponds to $0.2241/(2\pi)$ times f_s or, equivalently, $f_s/28.04$.

13.5.2 Stability of Discrete-Time Filters

To realize rational polynomials in z, discrete-time filters use delay elements (i.e., z^{-1} building blocks) much the same way that analog filters can be formed using integrators (i.e., s^{-1} building blocks). The result is that finite difference equations represent discrete-time filters rather than the differential equations used to describe continuous-time filters.

Consider the signal flow graph block diagram of a first-order, discrete-time filter shown in Fig. 13.12. A finite difference equation describing this block diagram can be written as

$$y(n+1) = bx(n) + ay(n) \tag{13.27}$$

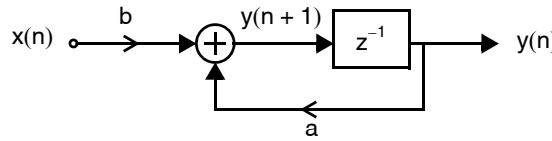


Fig. 13.12 A first-order, discrete-time filter.

In the z-domain, this equation is written as

$$zY(z) = bX(z) + aY(z) \tag{13.28}$$

where the z-domain property of delayed signals is used. We define the transfer function of this system to be $H(z)$ given by

$$H(z) \equiv \frac{Y(z)}{X(z)} = \frac{b}{z - a} \tag{13.29}$$

which has a pole on the real axis at $z = a$.

To test for stability, we let the input, $x(n)$, be an impulse signal (i.e., 1 for $n = 0$ and 0 otherwise), which gives the following output signal, according to (13.27),

$$y(0) = k$$

where k is some arbitrary initial state value for y.

$$\begin{aligned} y(1) &= b + ak \\ y(2) &= ab + a^2k \\ y(3) &= a^2b + a^3k \\ y(4) &= a^3b + a^4k \\ &\vdots \end{aligned}$$

More concisely, the response, $h(n)$, is given by

$$h(n) = \begin{cases} k & (n < 1) \\ (a^{n-1}b + a^n k) & (n \geq 1) \end{cases} \tag{13.30}$$

Clearly, this response remains bounded only when $|a| \leq 1$ for this first-order filter and is unbounded otherwise.

Although this stability result is shown only for first-order systems, in general, an arbitrary, linear, time-invariant, discrete-time filter, $H(z)$, is stable if and only if all its poles are located within the unit circle. In other words, if z_{pi} are the poles, then $|z_{pi}| < 1$ for all i. Locating some poles on the unit circle is similar to poles being on the imaginary jω-axis for continuous-time systems. For example, in the preceding first-order example, if $a = 1$, the pole is at $z = 1$, and the system is marginally stable (in fact, it is a discrete-time integrator). If we let $a = -1$, the pole is at $z = -1$, and one can show that the system oscillates at $f_s/2$, as expected.

Key Point: *A discrete-time filter is stable if and only if all its poles are located within the unit circle.*

13.5.3 IIR and FIR Filters

Key Point: *An infinite impulse response (IIR) filter has an impulse response that remains nonzero forever. A finite impulse response (FIR) filter has an impulse response that is finite in duration, eventually becoming precisely zero.*

Infinite impulse r esponse (IIR) filters are discrete-time filters that, when excited by an impulse, have nonzero outputs indefinitely, assuming infinite precision arithmetic.[2] For example, the filter given in (13.29) is an IIR filter (for a not equal to zero) since, although its impulse response decays towards zero (as all stable filters should), it remains nonzero forever, seen by letting $n \to \infty$ in (13.30).

Finite impulse response (FIR) filters are discrete-time filters that, when excited by an impulse, their outputs go precisely to zero (and remain zero) after a finite value of n. As an example of an FIR filter, consider the following filter:

$$y(n) = \frac{1}{3}[x(n) + x(n-1) + x(n-2)] \tag{13.31}$$

The transfer function for this filter, $H(z)$, is

$$H(z) = \frac{1}{3}\sum_{i=0}^{2} z^{-i} \tag{13.32}$$

This filter is essentially a running average filter since its output is equal to the average value of its input over the last three samples. Applying an impulse signal to this filter results in an output that is nonzero for only three samples and, therefore, this is an FIR filter. Note that this FIR filter has poles, but they all occur at $z = 0$.

Some advantages of FIR filters are that stability is never an issue (they are always stable) and exact linear phase filters can be realized (a topic beyond the scope of this chapter). However, for many specifications, an IIR filter can meet the same specifications as an FIR filter, but with a much lower order, particularly in narrowband filters in which the poles of an IIR filter are placed close to the unit circle (i.e., has a slowly decaying impulse response).

13.5.4 Bilinear Transform

One method for obtaining a discrete-time transfer functions that meets a set of specifications is to use a bilinear transform to convert the specifications into a set of fictitious continuous-time filter specifications, design a filter in continuous-time to meet those specifications, and then convert the resulting filter back to discrete-time, again using a bilinear transform. Assuming that $H_c(s)$ is a continuous-time transfer function (where s is the complex variable equal to $\sigma_s + j\Omega$), the bilinear transform is defined to be given by[3]

$$s = \frac{z-1}{z+1} \tag{13.33}$$

The inverse transformation is given by

$$z = \frac{1+s}{1-s} \tag{13.34}$$

2. The impulse response of an IIR filter may become so small that, with finite precision arithmetic, it is rounded to zero.
3. It should be noted here that in many other textbooks, the bilinear transform is defined as $s = (2/T)[(z-1)/(z+1)]$ where T is the sampling period. Here, we have normalized T to 2 since we use the bilinear transform only as a temporary transformation to a continuous-time equivalent, and then we inverse transform the result back to discrete time. Thus, the value of T can be chosen arbitrarily as long as the same value is used in each transformation.

A couple of points of interest about this bilinear transform are that the z-plane locations of 1 and –1 (i.e., dc and $f_s/2$) are mapped to s-plane locations of 0 and ∞, respectively. However, with a little analysis, we will see that this bilinear transformation also maps the unit circle, $z = e^{j\omega}$, in the z-plane to the entire $j\Omega$-axis in the s-plane. To see this mapping, we substitute $z = e^{j\omega}$ into (13.33),

> ***Key Point:*** *The bilinear transform provides a mapping between discrete- and continuous-time transfer functions whereby $z = -1$ maps to infinite frequency and $z = 1$ maps to dc.*

$$s = \frac{e^{j\omega} - 1}{e^{j\omega} + 1} = \frac{e^{j(\omega/2)}\left(e^{j(\omega/2)} - e^{-j(\omega/2)}\right)}{e^{j(\omega/2)}\left(e^{j(\omega/2)} + e^{-j(\omega/2)}\right)} \tag{13.35}$$

$$= \frac{2j\,\sin(\omega/2)}{2\,\cos(\omega/2)} = j\tan(\omega/2)$$

Thus, we see that points on the unit circle in the z-plane are mapped to locations on the $j\Omega$-axis in the s-plane, and we have

$$\Omega = \tan(\omega/2) \tag{13.36}$$

As a check, note that the z-plane locations of 1 and –1, which correspond to ω equal to 0 and π, respectively, map to Ω equal to 0 and ∞.

One way to use this transform is to design a continuous-time transfer function, $H_c(s)$, and choose the discrete-time transfer function, $H(z)$, such that

$$H(z) \equiv H_c[(z-1)/(z+1)] \tag{13.37}$$

> ***Key Point:*** *Using the bilinear transform, discrete-time transfer functions can be designed by mapping the specifications to fictitious continuous-time equivalents and finding a suitable $H_c(s)$. The inverse mapping $H(z) = H_c((z-1)/(z+1))$ results in the desired discrete-time transfer function.*

With such an arrangement, one can show that,

$$H(e^{j\omega}) = H_c[j\tan(\omega/2)] \tag{13.38}$$

and so the response of $H(z)$ is shown to be equal to the response of $H_c(s)$, except with a frequency warping according to (13.36). Note that the order of $H(z)$ equals that of $H_c(s)$ since, according to (13.37), each s term is replaced by another first-order function.

EXAMPLE 13.6

Using the bilinear transform, find the 3-dB frequency of a first-order discrete-time filter, $H(z)$, that has a pole at 0.8 and a zero at –1.

Solution

Using (13.33), we see that the z-plane pole of 0.8 and zero of –1 are mapped to a pole at –0.11111 and to a zero at ∞ in the s-plane, respectively. Such a continuous-time filter has a 3-dB frequency at $\Omega = 0.11111$ rad/s. Using (13.36), we find that the equivalent 3-dB frequency value in the z-plane is given by $\omega = 0.2213$ rad/sample, or $f_s/28.4$.

Note that this result is very close to that in Example 13.5. However, the two results are not in exact agreement since a zero at –1 is also present in this example.

EXAMPLE 13.7

Using the bilinear transform, find a first-order $H(z)$ that has a 3-dB frequency at $f_s/20$, a zero at -1, and a dc gain of 1.

Solution

Using (13.36), the frequency value, $f_s/20$, or equivalently, $\omega = (2\pi)/20 = 0.314159$ is mapped to $\Omega = 0.1584$ rad/s. Thus, $H_c(s)$ should have a 3-dB frequency value of 0.1584 rad/s. Such a 3-dB frequency value is obtained by having a s-plane zero equal to ∞ and a pole equal to -0.1584. Transforming these continuous-time pole and zero back into the z-plane using (13.34) results in a z-plane zero at -1 and a pole at 0.7265. Therefore, $H(z)$ appears as

$$H(z) = \frac{k(z+1)}{z - 0.7265}$$

The constant k can be determined by setting the dc gain to 1, or equivalently, $|H(1)| = 1$, which results in $k = 0.1368$.

13.6 SAMPLE-AND-HOLD RESPONSE

In this section, we look at the frequency response that occurs when we change a discrete-time signal back into an analog signal with the use of a sample-and-hold circuit. Note that here we plot a frequency response for all frequencies (as opposed to only up to $f_s/2$) since the output signal is a continuous-time signal rather than a discrete-time one.

A sample-and-hold signal, $x_{sh}(t)$, is related to its sampled signal by the mathematical relationship

$$x_{sh}(t) = \sum_{n=-\infty}^{\infty} x_c(nT)[\vartheta(t - nT) - \vartheta(t - nT - T)] \tag{13.39}$$

Note that, once again, $x_{sh}(t)$ is well defined for all time, and thus the Laplace transform can be found to be equal to

$$\begin{aligned} X_{sh}(s) &= \frac{1 - e^{-sT}}{s} \sum_{n=-\infty}^{\infty} x_c(nT)e^{-snT} \\ &= \frac{1 - e^{sT}}{s} X_s(s) \end{aligned} \tag{13.40}$$

This result implies that the hold transfer function, $H_{sh}(s)$, is equal to

$$H_{sh}(s) = \frac{1 - e^{-sT}}{s} \tag{13.41}$$

It should be mentioned here that this transfer function is usually referred to as the *sample-and-hold response* although, in fact, it only accounts for the hold portion.

The spectrum for $H_{sh}(s)$ is found by substituting $s = j\omega$ into (13.41), resulting in

$$H_{sh}(j\omega) = \frac{1 - e^{-j\omega T}}{j\omega} = T \times e^{-j\omega T/2} \times \frac{\sin(\omega T/2)}{(\omega T/2)} \tag{13.42}$$

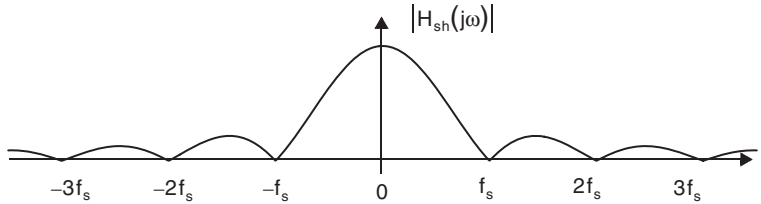


Fig. 13.13 Sample-and-hold response (also called the sinc response).

The magnitude of this response is given by

$$|H_{sh}(j\omega)| = T\frac{|\sin(\omega T/2)|}{|\omega T/2|}$$

or

$$|H_{sh}(f)| = T\frac{|\sin(\pi f/f_s)|}{|(\pi f)/f_s|} \tag{13.43}$$

and is often referred to as the $(\sin x)/x$ or *sinc response*. This magnitude response is illustrated in Fig. 13.13.

Multiplying this sinc response by $X_s(f)$ in Fig. 13.2 confirms the spectral relationship for $X_{sh}(f)$. It should be noted here that this frequency shaping of a sample and hold occurs only for a continuous-time signal. Specifically, although the sample and hold before the A/D converter shown in Fig. 13.1(*b*) would result in $x_{sh}(t)$ having smaller images at higher frequencies (due to the sinc response), the images of $x(n)$ are all of the same height (i.e., they are not multiplied by the sinc response) since it is a discrete-time signal. This is because, when converted to a discrete-time signal, the smaller images located at and above f_s are aliased in a manner that precisely reconstructs the original spectrum, so long as it was sampled above the Nyquist rate. In other words, a sample and hold before an A/D converter simply allows the converter to have a constant input value during one conversion and does not aid in any anti-aliasing requirement.

> ***Key Point:*** *The frequency-response of a sample-and-held signal is that of the original discrete-time sequence, repeated every f_s, and then shaped by a sinc frequency response.*

EXAMPLE 13.8

Consider the discrete-time signal processing system shown in Fig. 13.1(*b*), where a sample (and clock rate) of 50 kHz is used and the digital filter has the response

$$H(z) = \frac{0.2}{z - 0.8}$$

For a 10-kHz input sinusoidal signal of 1 V rms, find the magnitude of the output signal, $y_{sh}(t)$, at 10 kHz, and at the images 40 kHz and 60 kHz. Assume the quantization effects of the A/D and D/A converters are so small that they can be ignored. Also assume the converters complement each other such that a 1-V dc signal into the A/D converter results in a 1-V dc signal from the D/A converter if the converters are directly connected together.

Solution

First, the magnitude of $H(z)$ is found for a 10-kHz signal when the sampling rate is 50 kHz by noting that

$$\begin{aligned} e^{j(2\pi \times 10\text{ kHz}/50\text{ kHz})} &= e^{j0.4\pi} = \cos(0.4\pi) + j\sin(0.4\pi) \\ &= 0.309017 + j0.951057 \end{aligned}$$

and therefore the gain of $H(z)$ for a 10-kHz signal is found to be given by

$$|H(e^{j0.4\pi})| = \frac{0.2}{|(0.309017 - 0.8) + j0.951057|} = 0.186864$$

To determine the magnitude of $y_{sh}(t)$ at various frequencies, we need only multiply the filter's gain by the sinc response shown in Fig. 13.13. Specifically, the magnitude of the sample-and-hold response for a 10-kHz signal when using a 50-kHz clock is equal to

$$|H_{sh}(10\text{ kHz})| = T\frac{\sin(\pi 10/50)}{(\pi 10/50)} = 0.9355T$$

and therefore the magnitude of $y_{sh}(t)$ at 10 kHz is equal to $0.18686 \times 0.9355 = 175\ \text{mV}_{\text{rms}}$ (note that the T term cancels with a similar T term due to the creation of a discrete-time signal—see (13.13) in Section 13.2). Similarly, the magnitude of $H_{sh}(f)$ at 40 kHz and 60 kHz is 0.2399 and 0.1559, respectively, resulting in the magnitude of $y_{sh}(t)$ at 40 kHz and 60 kHz to be 43.7 and 29.1 mV_{rms}, respectively.

13.7 KEY POINTS

- Discrete-time signals may be processed either digitally, in which case they are quantized both in time and amplitude, or using discrete-time analog circuits such as switched-capacitor circuits, in which case the signal amplitude is continuously variable. [p. 537]
- The spectrum of a sampled signal is a sum of shifted copies of the original continuous-time spectrum so that no aliasing occurs if the original waveform is bandlimited to one-half the sampling rate. The minimum sampling frequency required to avoid aliasing is called the Nyquist rate, and is equal to two times the signal bandwidth. [p. 540]
- Discrete-time signal spectra are equal to the corresponding sampled signal spectra with the sample time normalized to 1 Hz. Hence, their spectra repeat every 2π. [p. 541]
- To avoid aliasing when downsampling by a factor L, the original signal must be sampled at L times the Nyquist rate. [p. 543]
- Upsampling by L results in a renormalization along the frequency axis so that L copies of the original signal spectrum are now squeezed in the range 0 to 2π. [p. 543]
- The frequency response of a discrete-time filter H(z) is found by evaluating its magnitude and phase on the unit circle, or graphically by considering the magnitude and phase of vectors connecting its poles and zeros to points around the unit circle. [p. 546]
- In discrete-time, the point z = 1 corresponds both to dc and to signals at the sampling frequency. The frequency response is periodic, and for rational transfer functions with real-valued coefficients exhibits conjugate symmetry around dc. Hence, discrete-time frequency responses need only be plotted from dc to half the sample rate, p. [p. 546]
- A discrete-time filter is stable if and only if all its poles are located within the unit circle. [p. 549]
- An infinite impulse response (IIR) filter has an impulse response that remains nonzero forever. A finite impulse response (FIR) filter has an impulse response that is finite in duration, eventually becoming precisely zero. [p. 550]
- The bilinear transform provides a mapping between discrete- and continuous-time transfer functions whereby z = – 1 maps to infinite frequency and z = 1 maps to dc. [p. 551]
- Using the bilinear transform, discrete-time transfer functions can be designed by mapping the specifications to fictitious continuous-time equivalents and finding a suitable $H_c(s)$. The inverse mapping $H(z) = H_c((z - 1)/(z + 1))$ results in the desired discrete-time transfer function. [p. 551]
- The frequency-response of a sample-and-held signal is that of the original discrete-time sequence, repeated every f_s, and then shaped by a sinc frequency response. [p. 553]

13.8 REFERENCES

R. Gregorian and G. C. Temes. *Analog MOS Integrated Circuits for Signal Processing*. John Wiley & Sons, New York, 1986.

A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey, 1989.

J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications*. Macmillan, New York, 1992.

P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

13.9 PROBLEMS

13.9.1 Section 13.1: Overview of Some Signal Spectra

13.1 In Fig. 13.1, $x_c(t) = 2\sin(2\pi \cdot 10\text{ kHz} \cdot t)$ and the A/D converter sampling rate is 50 kHz. Sketch time- and frequency-domain plots of x_c, x_s, x_{sh}, and x.

13.2 In Fig. 13.1, $y(n) = 3\sin((2\pi/10) \cdot n)$ and the D/A converter with hold is clocked at 5 MHz. Sketch time- and frequency-domain plots of y, y_{sh}, and y_c.

13.9.2 Section 13.2: Laplace Transforms of Discrete-Time Signals

13.3 A 51 MHz sinusoid is sampled at a frequency f_s by the periodic pulse train $s(t)$ in (13.9) resulting in a continuous-time signal $x_s(t)$. Sketch the spectrum of the resulting sampled signal $X_s(f)$ when

a. $f_s = 200$ MHz

b. $f_s = 100$ MHz

c. $f_s = 50$ MHz

d. $f_s = 2$ MHz

13.4 What is the minimum sampling frequency, f_s required to sample the following signal without aliasing: $x_c(t) = \sin(2\pi \cdot 50\text{ kHz} \cdot t) + 3\cos(2\pi \cdot 150\text{ kHz} \cdot t)$?

13.9.3 Section 13.3: z-Transform

13.5 Show that if $x(n) \leftrightarrow X(z)$, then $x(n-k) \leftrightarrow z^{-k}X(z)$.

13.6 Find an expression for the discrete-time spectrum $X(\omega)$ obtained by sampling a square pulse of unit amplitude and 1 ms in duration at a sampling rate of 10 kHz.

13.7 If a discrete-time sequence with a frequency of $\pi/10$ was obtained at a sampling rate of 10 MHz, what was the frequency of the original continuous-time signal (assuming no aliasing occurred)?

13.8 A discrete-time sequence $y(n)$ is obtained from $x(n) \leftrightarrow X(z)$ as follows:

$$y(n) = x(n) + 2x(n-3)$$

What is the resulting z-transform, $Y(z)$?

13.9.4 Section 13.4: Downsampling and Upsampling

13.9 What is the maximum amount of downsampling that can be applied to $x_{s2}(n)$ in Example 13.1 before aliasing occurs?

13.10 Is downsampling a time-invariant operation? If so, explain. If not, give a simple counter example. Is downsampling a linear operation?

13.11 Repeat Example 13.3, but upsample by 8 (i.e., insert 7 zeros between samples).

13.12 Upsampling and downsampling are often used for converting between two sample rates. Give a block diagram showing how one might convert 0–20 kHz band-limited audio signal from a sampling rate of 50 kHz to 40 kHz.

13.9.5 Section 13.5: Discrete-Time Filters

13.13 Find the location of the poles and zeros of the transfer function given by $H(z)$ in (13.21).

13.14 Find the location of the pole required in a first-order filter that has its zero at ∞ such that its 3-dB frequency is at $f_s/10$.

13.15 Sketch the magnitude response (in dB) of the following transfer function, $H(z)$, by finding the gains at dc and at $\omega = \pi$, as well as the location of the 3-dB frequency.

$$H(z) = \frac{0.05z}{z - 0.95}$$

13.16 Consider the following difference equation, which describes a filter with input $u(n)$ and output $y(n)$.

$$y(n) = u(n) - 0.3y(n-1)$$

Determine $y(n)$ when $u(n)$ is an impulse (i.e., when the impulse is 1 at $n = 1$ and 0 otherwise) assuming $y(0) = 0$.

13.17 Find a first-order $H(z)$ that has a 3-dB frequency at $f_s/30$, a zero at –1, and a dc gain of 1. Sketch the magnitude response from $\omega = 0$ to $\omega = 2\pi$.

13.18 Find the coefficients a and b in Fig. 13.12 required to realize a lowpass filter with a 3-dB frequency of $f_s/50$ and a dc gain of 2.

13.19 Using a bilinear transform, find an $H(z)$ to realize a second-order transfer function with a dc gain of one, a maximally-flat passband (i.e., $Q = 1/\sqrt{2}$ in the continuous-time domain), and a –3-dB frequency at 1 kHz when the sampling rate is 100 kHz.

13.20 Find the output sequence of a discrete-time integrator, $H(z) = 1/(z-1)$, when its input is a sinusoidal waveform of peak height 1 at $f_s/4$ and its initial state is zero. In other words, the input waveform is the series of numbers $\{1, 0, -1, 0, 1, 0, -1, \ldots\}$. Using a frequency analysis, what is the magnitude and phase response for $H(z)$ at $f_s/4$? Check the frequency domain result with the time domain result.

13.21 A 0–5-kHz voiceband microphone is modelled as having a low-pass, first-order frequency response with a pole at 10 kHz. If the microphone output is sampled at 100 kHz, how much attenuation is required in the anti-aliasing filter for 80-dB noise rejection around 100 kHz?

13.9.6 Section 13.6: Sample-and-Hold Response

13.22 A 100-Hz sinusoidal signal, $x(n)$, is digitally created using a 400-Hz sampling rate producing a sample-and-hold signal, $x_{sh}(t)$. Before any smoothing filter is applied and assuming the magnitude of $x_{sh}(t)$ at 100 Hz is 1 V_{rms}, what is the magnitude of $x_{sh}(t)$ at 300 Hz? Also, find the magnitude of $x_{sh}(t)$ at 500 Hz and at 700 Hz.

13.23 Repeat Problem 13.22, but assume the 100-Hz signal is digitally created using a 10-kHz signal, and find the magnitude of $x_{sh}(t)$ at 9.9 kHz and at 10.1 kHz.

13.24 Repeat Example 13.8, but assume an input signal frequency of 100 Hz and find the magnitude of the output signal, $y_{sh}(t)$, at 100 Hz, 99.9 kHz, and 100.1 kHz.

13.25 A particular D/A converter is clocked at 100 MHz and produces an output that looks like $\tau x_s(t)$ in Fig. 13.3 with finite τ. Find the value of τ required so that, when producing a 5 MHz sinusoid, the amplitudes of the images are one-quarter of the fundamental signal amplitude.

CHAPTER

14 Switched-Capacitor Circuits

Perhaps the most popular approach for realizing analog signal processing in MOS (or BiCMOS) integrated circuits is through the use of switched-capacitor circuits [Poschenrieder, 1966; Fried, 1972; Caves, 1977; Hosticka, 1977; Young, 1977]. A switched-capacitor circuit operates as a discrete-time signal processor (although without the use of A/D or D/A converters). As a result, these circuits are most easily analyzed with the use of z-transform techniques and typically require anti-aliasing and smoothing filters when combined with continuous-time circuits.

As a filtering technique, switched-capacitor circuits are popular due to their accurate frequency response as well as good linearity and dynamic range. Accurate discrete-time frequency responses are obtained since filter coefficients are determined by capacitance ratios which can be set quite precisely in an integrated circuit (on the order of 0.1 percent). Such an accuracy is orders of magnitude better than that which occurs for integrated RC time constants (which can vary by as much as 20 percent). Once the coefficients of a switched-capacitor discrete-time filter are accurately determined, its overall frequency response remains a function of the clock (or sampling) frequency. Fortunately, clock frequencies can also be set very precisely through the use of a crystal oscillator.

In addition to creating filtering functions, switched-capacitor circuit techniques can be used to realize a variety of other signal-processing blocks such as gain-stages, voltage-controlled oscillators, and modulators. This chapter will introduce the basic principles of switched-capacitor filters as well as some other nonfiltering functions realized using switched-capacitor circuits.

14.1 BASIC BUILDING BLOCKS

A switched-capacitor circuit is realized with the use of some basic building blocks such as opamps, capacitors, switches, and nonoverlapping clocks. Brief descriptions of each of these blocks and their important nonidealities with respect to switched-capacitor circuits are given in this section.

14.1.1 Opamps

The basic principles of switched-capacitor circuits can be well understood assuming ideal opamps. However, some important opamp nonidealities in practical switched-capacitor circuits are dc gain, unity-gain frequency and phase-margin, slew-rate, and dc offset. Opamp input impedance is generally capacitive assuming MOSFET input stages are used.[1]

Key Point: *Opamp nonidealities such as dc gain, unity-gain frequency and phase-margin, slew-rate, and dc offset impact switched-capacitor circuit performance.*

The *dc gain* of opamps in a MOS technology intended for switched-capacitor circuits is typically on the order of 40 to 80 dB. Low dc gains affect the coefficient accuracy of the discrete-time transfer function of a switched-capacitor filter.

1. BJT input stages are rarely used as they would result in charge leakage due to their nonzero input bias current.

The *unity-gain frequency and phase margin* of an opamp gives an indication of its small signal settling behavior. A general rule of thumb is that the clock frequency of a switched-capacitor circuit should be at least five times lower than the unity-gain frequency of the opamp, assuming little slew-rate behavior occurs and the phase margin is greater than 70 degrees. Modern SC circuits are often realized using high-frequency single-stage opamps having very large output impedances (up to 100 kΩ or much larger). Since the loads of these opamps are purely capacitive (never resistive), their dc gains remain high despite the lack of an output buffer stage. It should be noted here that their unity-gain frequency and phase margin are determined by the load capacitance, which also serves as the compensation capacitor. Thus, in these single-stage opamps, doubling the load capacitance would halve the unity-gain frequency and *improve* the phase margin.

The finite *slew rate* of an opamp can limit the upper clock rate in a switched-capacitor circuit as these circuits rely on charge being quickly transferred from one capacitor to another. Thus, at the instance of charge transfer, it is not uncommon for opamps to slew-rate limit.

A nonzero *dc offset* can result in a high output dc offset for the circuit depending on the topology chosen. Fortunately, a technique known as correlated double sampling can significantly reduce this output offset and at the same time reduce low-frequency opamp input noise (i.e., 1/f noise).

14.1.2 Capacitors

A highly linear capacitance in an integrated circuit is typically constructed from two closely-spaced conducting layers, as shown in Fig. 14.1(*a*).[2] The desired capacitance, C, is formed as the intersection of area between the two conductive layers, layer 1 and layer 2. However, since the substrate below layer 1 is an ac ground (the substrate is connected to one of the power supplies or analog ground), there also exists a substantial parasitic capacitance, C_{p1}, which may be as large as 20 percent of C_1. This large parasitic capacitance, C_{p1}, is known as the bottom plate capacitance. A top plate capacitance, C_{p2}, also exists due primarily to the interconnect capacitance, but it is typically much smaller (generally less than 5 percent of C). In summary, the equivalent model for a single integrated capacitor is three capacitors, as shown in Fig. 14.1(*b*), where the bottom plate is often explicitly shown to indicate that a larger parasitic capacitance occurs on that node.

14.1.3 Switches

The requirements for switches used in switched-capacitor circuits are that they have a very high off resistance (so little charge leakage occurs), a relatively low on resistance (so that the circuit can settle in less than half the clock

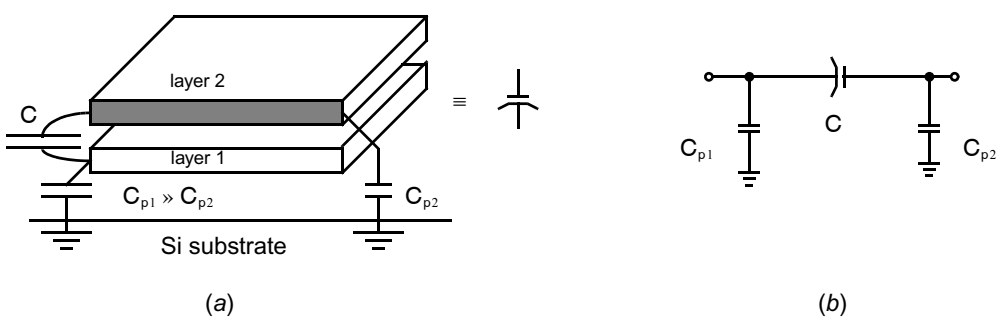


Fig. 14.1 An integrated circuit capacitor for switched-capacitor circuits: (*a*) physical construction; (*b*) circuit model.

2. These may be two layers of polysilicon with a specially-constructed, very thin dielectric layer separating them, thus providing a high capacitance per unit area, or the layers might be polysilicon and heavily-doped silicon, or two metal layers.

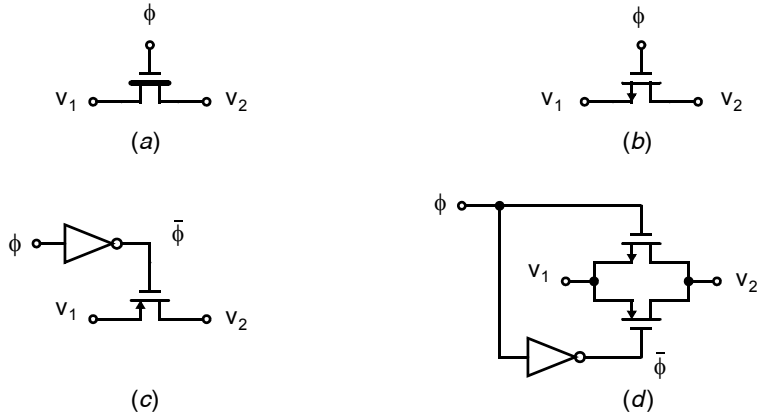


Fig. 14.2 Switch symbol and some transistor circuits: (*a*) symbol, (*b*) n-channel switch, (*c*) p- channe switch, (*d*) transmission gate.

period), and introduce no offset voltage when turned on (as does a bipolar switch whose on voltage equals $V_{CE(sat)}$). The use of MOSFET transistors as switches satisfies these requirements, as MOSFET switches can have off resistances up to the GΩ range, have no offset on voltages, and have on resistances of 5 kΩ or much less depending on transistor sizing.

The symbol for a switch and some MOSFET circuits that realize a switch are shown in Fig. 14.2. Here, the signal ϕ is one of two logic levels typically corresponding to the maximum and minimum power supply levels. As a convention, when the clock signal ϕ is high, the switch is assumed to be on (i.e., shorted). Although a switch can be implemented using a single transistor, as shown in Fig. 14.2(*b*) and Fig. 14.2(*c*), the signal range of the switch is reduced. For example, consider the ideal n-channel transistor switch shown in Fig. 14.2(*b*) being used in a circuit having power supplies of 0 to 1.8 V and a threshold voltage of $V_{tn} = 0.45$ V. The switch can always be turned off irrespective of the input voltage by putting 0 V on its gate. However, when the switch is on, its gate voltage will be $V_{DD} = 1.8$ V; in this case, the switch can only equalize the two voltages if V_1 and/or V_2 are lower than $V_{DD} - V_{tn}$ or around 1.3 V taking the body effect into account. Thus, the signal range for this switch is limited from 0 to 1.3 V. A similar argument can be made for the p-channel switch shown in Fig. 14.2(*c*) except that its signal range would be limited from 0.5 to 1.8 V. While this signal range is acceptable in some situations, the full signal range of 0 to 1.8 V can be achieved using two transistors in parallel, as shown in Fig. 14.2(*d*). This combination is commonly called either a CMOS switch (as opposed to the NMOS switch of Fig. 14.2(*a*)) or a CMOS transmission gate.

Some of the important nonideal switch effects in switched-capacitor circuits are the nonlinear junction capacitance on each side of the switch, channel charge injection, and capacitive coupling from the logic signal, ϕ, to each side of the switch. Nonlinear capacitance effects are alleviated through the use of parasitic insensitive structures, as discussed in Section 14.2, while channel charge injection is reduced by careful switch timing, as seen in Section 14.6.

14.1.4 Nonoverlapping Clocks

At least one pair of nonoverlapping clocks is essential in switched-capacitor circuits. These clocks determine when charge transfers occur and they must be nonoverlapping in order to guarantee charge is not inadvertently lost. As seen in Fig. 14.3(*a*), the term *nonoverlapping clocks* refers to two logic signals running at the same frequency and arranged in such a way that at no time are both signals high. Note that the time axis in Fig. 14.3(*a*) has been normalized with respect to the clock period, T. Such a normalization illustrates the location of the sample numbers of the discrete-time signals that occur in switched-capacitor filters. As a convention, we denote

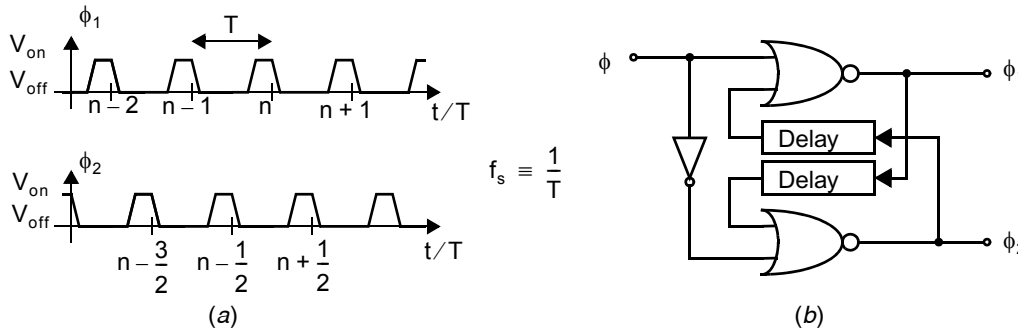


Fig. 14.3 Nonoverlapping clocks. (*a*) Clock signals, ϕ_1 and ϕ_2. (*b*) A possible circuit implementation of nonoverlapping clocks from a single clock.

the sampling numbers to be integer values (i.e., $(n-1)$, (n), $(n+1)$, etc.) just before the end of clock phase ϕ_1, while the end of clock phase ϕ_2 is deemed to be $1/2$ sample off the integer values as shown (i.e., $(n-3/2)$, $(n-1/2)$, etc.). However, it should be noted that it is not important that the falling clock edge of ϕ_2 occur precisely one-half a clock period earlier than the falling edge of ϕ_1. In general, the locations of the clock edges of ϕ_1 and ϕ_2 need only be moderately controlled to allow for complete charge settling (assuming separate low-jitter sample and holds are used at the input and output of the overall circuit).

One simple method for generating nonoverlapping clocks is shown in Fig. 14.3(*b*) [Martin, 1981]. Here, delay blocks are used to ensure that the clocks remain nonoverlapping. These delays could be implemented as a cascade of an even number of inverters or perhaps an RC network.

14.2 BASIC OPERATION AND ANALYSIS

14.2.1 Resistor Equivalence of a Switched Capacitor

Consider the switched-capacitor circuit shown in Fig. 14.4(*a*), where V_1 and V_2 are two dc voltage sources. To analyze this circuit's behavior, we analyze the circuit from a charge perspective. Recall that the charge on a capacitor, Q_x, is equal to the capacitance value, C_x, times the voltage across it, V_x. In mathematical terms, we have

$$Q_x = C_x V_x \tag{14.1}$$

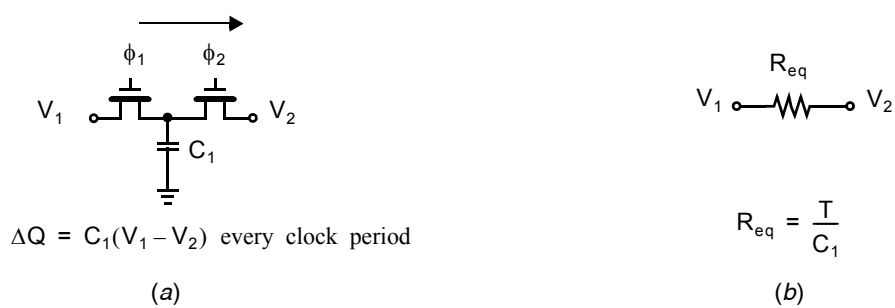


Fig. 14.4 Resistor equivalence of a switched capacitor. (*a*) Switched-capacitor circuit, and (*b*) resistor equivalent.

Now, since ϕ_1 and ϕ_2 are assumed to be a pair of nonoverlapping clocks, C_1 is charged to V_1 and then V_2 during each clock period. Therefore, the change in charge (in coulombs) over one clock period, ΔQ_1, that is transferred from node V_1 into node V_2 is given by

$$\Delta Q_1 = C_1(V_1 - V_2) \tag{14.2}$$

Since this charge transfer is repeated every clock period, we can find the equivalent average current due to this charge transfer by dividing by the clock period. In other words, the average current, I_{avg} (in amps), is equal to the average number of coulombs/second and is given by

$$I_{avg} = \frac{C_1(V_1 - V_2)}{T} \tag{14.3}$$

where T is the clock period (i.e., the sampling frequency $f_s = 1/T$). Relating (14.3) to an equivalent resistor circuit shown in Fig. 14.4(*b*) where

$$I_{eq} = \frac{V_1 - V_2}{R_{eq}} \tag{14.4}$$

we see that the average current through the switched-capacitor circuit of Fig. 14.4(*a*) will equal that for the resistor circuit of Fig. 14.4(*b*) if

$$R_{eq} = \frac{T}{C_1} = \frac{1}{C_1 f_s} \tag{14.5}$$

This relation is intuitively satisfying because as the clock frequency is increased, the same charge is transferred each period but at a faster rate, so the average current is higher. If C_1 is increased, a larger amount of charge is transferred each period, which would also increase the average current. In addition, increasing the average current is equivalent to decreasing the equivalent resistance seen between the two voltage nodes, and therefore the equivalent resistance is inversely related to the capacitance and clock frequency.

Key Point: *When signal voltages change very slowly compared with the clock frequency, a switched-capacitor gives rise to an average current inversely proportional to the voltage across it, behaviorally equivalent to a resistor.*

Finally, it should be noted that this resistor approximation assumes that the charge transferred per clock cycle is constant over many cycles allowing one to arrive at an average current flow. Hence, it is mainly useful for looking at the low-frequency behavior of a switched-capacitor circuit, when voltages are changing very slowing compared to the clock frequency. For moderate input frequencies (in relation to the sampling frequency), an accurate discrete-time analysis is required, as discussed in the next subsection. Discrete-time analysis can be used in switched-capacitor circuits because the charge transfers are dependent on values of node voltages at particular instances in time and are ideally independent of transients during nonsampled times.

Key Point: *When signal voltages change appreciably from one clock cycle to the next, a discrete-time analysis is needed to accurately predict the input–output behavior of switched-capacitor circuits.*

EXAMPLE 14.1

What is the equivalent resistance of a 5 pF capacitance sampled at a clock frequency of 100 kHz?

Solution

Using (14.5), we have

$$R_{eq} = \frac{1}{(5 \times 10^{-12})(100 \times 10^{3})} = 2\ \text{M}\Omega$$

What is interesting here is that a very large impedance of 2 MΩ can be realized on an integrated circuit through the use of two transistors, a clock, and a relatively small capacitance. Such a large impedance typically requires a large amount of CMOS silicon area if it is realized as a resistor without any special processing fabrication steps.

14.2.2 Parasitic-Sensitive Integrator

An example of one of the first switched-capacitor discrete-time integrators [Hosticka, 1977] is shown in Fig. 14.5. Here, an extra switch is shown near the output voltage, $v_{co}(t)$, to indicate that the output signal is valid at the end of ϕ_1 in keeping with the sampling-time convention shown in Fig. 14.3(*a*). In other words, any circuit that uses the output of this discrete-time integrator should sample $v_{co}(t)$ on ϕ_1. Note however that no assumptions are made about when $v_{ci}(t)$ changes its value (it could be either ϕ_1 or ϕ_2); it is not important because $v_{ci}(t)$ is sampled at the end of ϕ_1 in the circuit shown.

Key Point: *Each clock cycle, a switched-capacitor integrator samples a quantity of charge proportional to an inputvoltage and the sampling capacitance, then delivers it to a second integrating capacitor.*

To analyze this circuit, we again look at the charge behavior and note that a virtual ground appears at the opamp's negative input. Assuming an initial integrator output voltage of $v_{co}(nT - T)$ implies that the charge on C_2 is equal to $C_2 v_{co}(nT - T)$ at time $(nT - T)$. Now also at time $(nT - T)$, ϕ_1 is just turning off (and ϕ_2 is off) so the input signal $v_{ci}(t)$ is sampled, resulting in the charge on C_1 being equal to $C_1 v_{ci}(nT - T)$, as shown in Fig. 14.6(*a*). When ϕ_2 goes high, its switch forces C_1 to discharge since it places a virtual ground on the top plate of C_1, as shown in Fig. 14.6(*b*). However, this discharging current must pass through C_2 and hence the charge on C_1 is added to the charge already present on C_2. Note that a positive

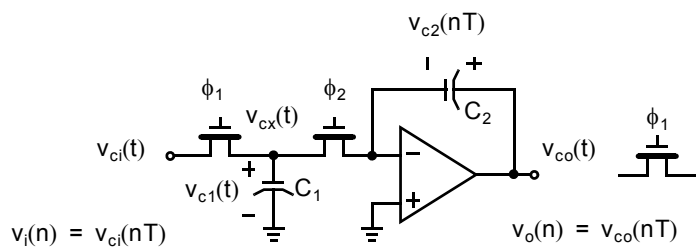


Fig. 14.5 A discrete-time integrator. This structure is sensitive to parasitic capacitances (not shown).

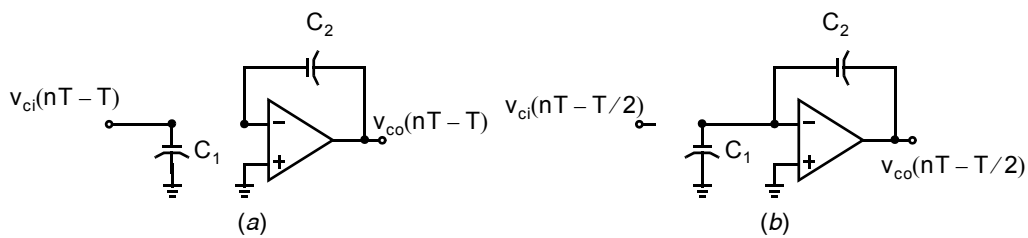


Fig. 14.6 The parasitic-sensitive integrator circuit for the two clock phases: (*a*) ϕ_1 (*b*) ϕ_2

input voltage will result in a negative voltage across C_2 and therefore a negative output voltage; thus, the integrator is an inverting integrator. At the end of ϕ_2 (once things have settled out), we can write the charge equation,

$$C_2 v_{co}(nT - T/2) = C_2 v_{co}(nT - T) - C_1 v_{ci}(nT - T). \tag{14.6}$$

The negative sign in (14.6) reflects the fact that the integrator is an inverting integrator.

Although (14.6) is correct, we would like to find the charge on C_2 at the end of ϕ_1 (or equivalently, nT) as indicated by the switch shown near $v_{co}(t)$ in Fig. 14.5. To find this charge, we note that once ϕ_2 turns off, the charge on C_2 will remain the same during the next ϕ_1, until ϕ_2 turns on again in its next cycle. Therefore, the charge on C_2 at time (nT) at the end of the next ϕ_1 is equal to that at time $(nT - T/2)$, or mathematically,

$$C_2 v_{co}(nT) = C_2 v_{co}(nT - T/2) \tag{14.7}$$

and we can combine (14.6) and (14.7) to write

$$C_2 v_{co}(nT) = C_2 v_{co}(nT - T) - C_1 v_{ci}(nT - T) \tag{14.8}$$

Dividing by C_2 and using the discrete-time variables $v_i(n) = v_{ci}(nT)$ and $v_o(n) = v_{co}(nT)$, we have the following discrete-time relationship for the circuit of Fig. 14.5:

$$v_o(n) = v_o(n-1) - \frac{C_1}{C_2} v_i(n-1) \tag{14.9}$$

Taking the z-transform of (14.9), we also have

$$V_o(z) = z^{-1} V_o(z) - \frac{C_1}{C_2} z^{-1} V_i(z) \tag{14.10}$$

from which we can finally find the transfer function, $H(z)$, to be given by

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\left(\frac{C_1}{C_2}\right) \frac{z^{-1}}{1 - z^{-1}} \tag{14.11}$$

Note that this transfer function realizes its gain coefficient as a ratio of two capacitances, and thus the transfer function can be accurately defined in an integrated circuit. Often, the transfer function would be rewritten to eliminate terms of z having negative powers to give

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\left(\frac{C_1}{C_2}\right) \frac{1}{z - 1} \tag{14.12}$$

Some typical voltage waveforms for this discrete-time integrator are shown in Fig. 14.7. Note here that the discrete-time equation derived only represents the relationship of the voltage signals $v_{ci}(t)$ and $v_{co}(t)$ at the time (nT), which occurs just before the end of ϕ_1. It does not say anything about what the voltages are at other times. In actual fact, the voltages are constant at other times (ignoring second-order effects), which is equivalent to taking the sampled voltages and then using them as inputs to a sample and hold for this particular SC integrator.

Key Point: *The discrete-time transfer function of a switched-capacitor integrator has a pole at dc, z = 1.*

EXAMPLE 14.2

Show that for frequencies much less than the sampling frequency, equation (14.11) approximates the transfer function of an ideal continuous-time integrator.

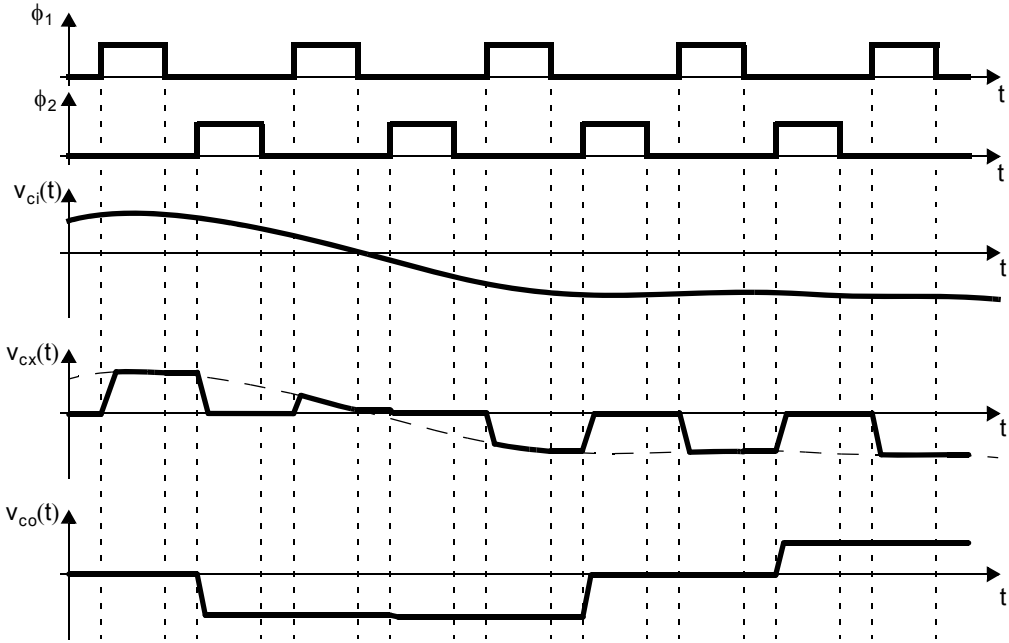


Fig. 14.7 Typical voltage waveforms for the discrete-time integrator shown in Fig. 14.5.

Solution

Equation (14.11) can be rewritten as

$$H(z) = -\left(\frac{C_1}{C_2}\right)\frac{z^{-1/2}}{z^{1/2} - z^{-1/2}} \tag{14.13}$$

To find the frequency response, we make use of the fact that

$$z = e^{j\omega T} = \cos(\omega T) + j\sin(\omega T) \tag{14.14}$$

where T, the period, is one over the sampling frequency. Therefore,

$$z^{1/2} = \cos\left(\frac{\omega T}{2}\right) + j\sin\left(\frac{\omega T}{2}\right) \tag{14.15}$$

and

$$z^{-1/2} = \cos\left(\frac{\omega T}{2}\right) - j\sin\left(\frac{\omega T}{2}\right) \tag{14.16}$$

Substituting (14.15) and (14.16) into (14.13) gives

$$H(z) = -\left(\frac{C_1}{C_2}\right)\frac{z^{-1/2}}{j2\sin\left(\frac{\omega T}{2}\right)} \cong -\left(\frac{C_1}{C_2}\right)\frac{z^{-1/2}}{j\omega T} \tag{14.17}$$

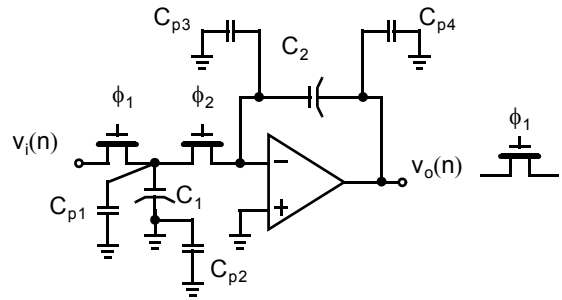


Fig. 14.8 A discrete-time integrator with parasitic capacitances shown.

for $\omega T << 1$. The $z^{-1/2}$ in the numerator represents a simple delay and can be ignored. Thus, we see that the transfer function is approximately that of a continuous-time integrator having a gain constant of

$$K_I \cong -\left(\frac{C_1}{C_2 T}\right) \tag{14.18}$$

which is a function of the integrator capacitor ratio and clock frequency only.

Thus far we have ignored the effect of the parasitic capacitances due to the creation of C_1 and C_2 as well as the nonlinear capacitances associated with the switches. The addition of such parasitic capacitances results in the circuit shown in Fig. 14.8. Here, C_{p1} represents the parasitic capacitance of the top plate of C_1 as well as the nonlinear capacitances associated with the two switches; C_{p2} represents the parasitic capacitance of the bottom plate of C_1; while C_{p3} represents the parasitic capacitance associated with the top plate of C_2, the input capacitance of the opamp, and that of the ϕ_2 switch. Finally, C_{p4} accounts for the bottom-plate parasitic capacitance of C_2 as well as any extra capacitance that the opamp output must drive.

In accounting for these parasitic capacitances, we can immediately discard the effect of C_{p2} since it always remains connected to ground. Similarly, the effect of C_{p3} on the transfer function is small since it is always connected to the virtual ground of the opamp. Also, C_{p4} is connected to the opamp output; therefore, although it may affect the speed of the opamp, it would not affect the final settling point of the opamp output. Finally, we note that C_{p1} is in parallel with C_1 and therefore the transfer function of this discrete-time integrator including the effects of parasitic capacitance is given by

$$H(z) = -\left(\frac{C_1 + C_{p1}}{C_2}\right)\frac{1}{z-1} \tag{14.19}$$

From (14.19), we see that the gain coefficient is related to the parasitic coefficient C_{p1}, which is not well controlled and would be partially nonlinear due to the input capacitances of the switches and top-plate parasitic capacitance. To overcome this deficiency, ingenious circuits known as parasitic-insensitive structures were developed [Martin, 1978; Allstot, 1978; Jacobs, 1978].

14.2.3 Parasitic-Insensitive Integrators

The parasitic-insensitive integrator was a critical development that allowed the realization of high-accuracy integrated circuits. In the first integrated integrator-based filters, the noninverting integrators (Fig. 14.9) were parasitic insensitive, whereas the inverting integrators (Fig. 14.8) were parasitic sensitive [Allstot, 1978]. The

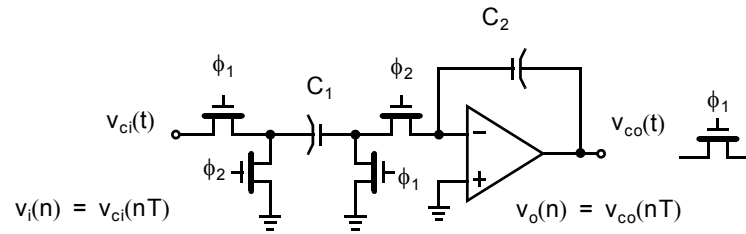


Fig. 14.9 A noninverting delaying discrete-time integrator that is not sensitive to parasitic capacitances.

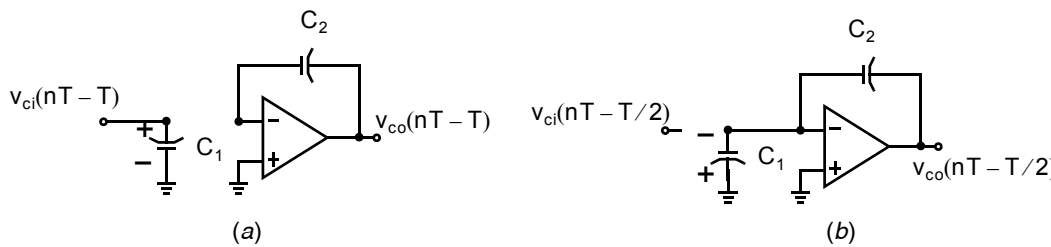


Fig. 14.10 The noninverting discrete-time integrator on the two clock phases: (*a*) ϕ_1 (*b*) ϕ_2

development of the parasitic-insensitive inverting integrator [Martin, 1978; Jacobs, 1978] allowed the complete filter to be insensitive, which greatly decreased second-order errors. To realize a parasitic-insensitive discrete-time integrator, two extra switches are used, as shown in Fig. 14.9. Analyzing this circuit as before, note that when ϕ_1 is on, C_1 is charged to $C_1 v_{ci}(nT - T)$, as shown in Fig. 14.10(*a*). When ϕ_2 turns on, C_1 is effectively "hooked" up backwards, and the discharge now occurs through what was the ground node rather than the input signal side, as shown in Fig. 14.10(*b*). Such a reverse connection results in $v_{co}(t)$ rising for a positive $v_{ci}(nT - T)$, and therefore the integrator is noninverting. With the rest of the analysis being performed as before, the transfer function, $H(z)$, for this discrete-time integrator is given as

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = \left(\frac{C_1}{C_2}\right) \frac{z^{-1}}{1 - z^{-1}} \tag{14.20}$$

Note that (14.20) represents a positive discrete-time integrator with a full delay from input to output (due to the z^{-1} in the numerator, which represents a one-period delay) so we refer to the circuit of Fig. 14.9 as a delaying discrete-time integrator. As was done for the parasitic sensitive inverting integrator of Fig. 14.8, one often rewrites the transfer function to eliminate negative powers of z yielding

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = \left(\frac{C_1}{C_2}\right) \frac{1}{z - 1} \tag{14.21}$$

Key Point: *In order to make a switched-capacitor integrator insensitive to the circuit's parasitic capacitances to ground, it is necessary to add two extra switches, for a total of four, on each switched capacitor.*

To investigate the behavior of this noninverting integrator with respect to parasitic capacitances, consider the same circuit drawn with parasitic capacitances, as shown in Fig. 14.11. Here, C_{p3} and C_{p4} do not affect the operation of the circuit as before. In addition, C_{p2} is either connected to ground through the ϕ_1 switch or to virtual ground through the ϕ_2 switch. Therefore, since C_{p2} always remains discharged (after settling), it also does not affect the operation of the circuit. Finally, C_{p1} is continuously being charged to $v_i(n)$ and discharged to ground. However during ϕ_1 on, the fact that C_{p1} is also charged to $v_i(n-1)$ does not affect the charge that is placed on C_1. When ϕ_2 is turned on, C_{p1} is discharged through the ϕ_2 switch attached to its node and none of its discharging current passes through C_1 to affect the charge accumulating on C_2. Therefore, it also does not affect the circuit operation. In summary, while the parasitic capacitances may slow down settling time behavior, they do not affect the discrete-time difference equation that occurs in the integrator shown in Fig. 14.9.

To obtain an inverting discrete-time integrator that is also parasitic insensitive, the same circuit as the parasitic-insensitive noninverting integrator can be used, but with the two switch phases on the switches near the opamp input side of C_1 (that is, the top-plate of C_1) interchanged as shown in Fig. 14.12 [Martin, 1978]. With this arrangement, note that the charge on C_2 does not change when ϕ_2 turns on (and ϕ_1 is off) as illustrated in Fig. 14.13(*b*). Therefore, we can write

$$C_2 v_{co}(nT - T/2) = C_2 v_{co}(nT - T) \tag{14.22}$$

Also note that C_1 is fully discharged at the end of ϕ_2 on. Now when ϕ_1 is turned on in Fig. 14.13(*a*), C_1 is charged up to $v_{ci}(t)$. However, the current needed to charge C_1 passes through C_2, in effect changing the charge on C_2 by that amount. At the end of ϕ_1 on, the charge left on C_2 is equal to its old charge (before ϕ_1 was on)

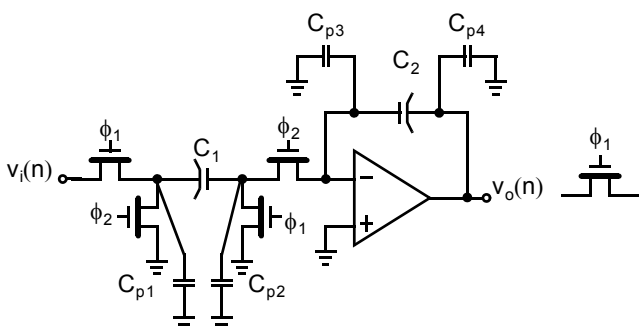


Fig. 14.11 A parasitic-insensitive delayed integrator with parasitic capacitances shown.

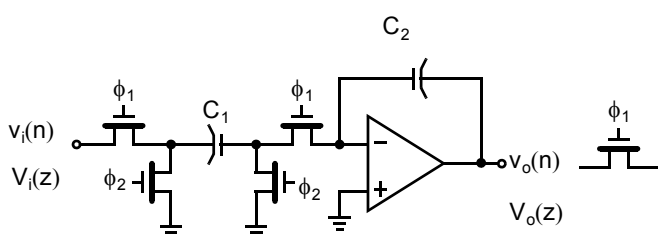


Fig. 14.12 A delay-free discrete-time integrator (not sensitive to parasitic capacitances).

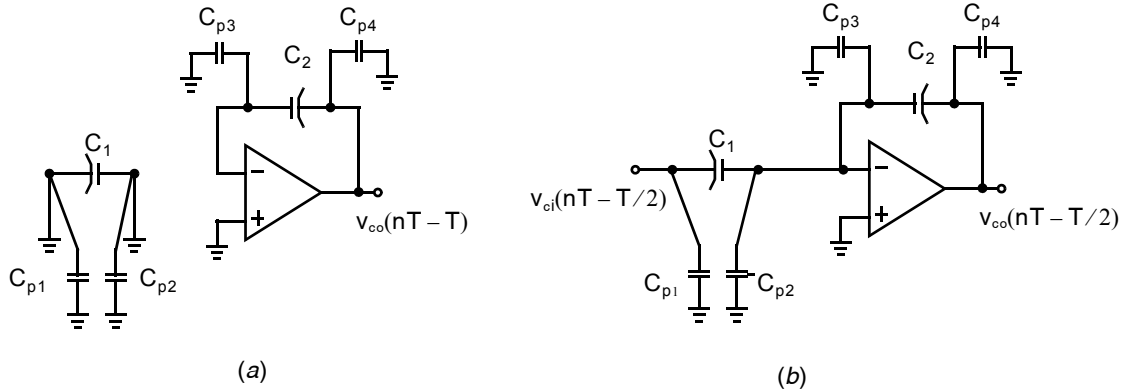


Fig. 14.13 A parasitic-insensitive delay-free integrator on the two clock phases: (*a*) ϕ_2 (*b*) ϕ_1

subtracted from the charge needed to charge up C_1 to $v_{ci}(nT)$. Thus, we have the following charge relationship:

$$C_2 v_{co}(nT) = C_2 v_{co}(nT - T/2) - C_1 v_{ci}(nT) \tag{14.23}$$

Note that $v_{ci}(nT)$ occurs in the preceding equation rather than $v_{ci}(nT - T)$, since the charge on C_2 at the end of ϕ_1 is related to the value of $v_{ci}(nT)$ at the same time, and therefore the integrator is considered to be delay free. Substituting in (14.22), dividing by C_2, and switching to discrete-time variables, we have

$$v_o(n) = v_o(n-1) - \frac{C_1}{C_2} v_i(n) \tag{14.24}$$

Taking the z-transform of both sides, the transfer function, $H(z)$, for this delay-free integrator is found to be

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\left(\frac{C_1}{C_2}\right) \frac{1}{1 - z^{-1}} \tag{14.25}$$

where the fact that the numerator has a purely real term is indicative of the integrator being delay free. As before, (14.25) is often rewritten as

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\left(\frac{C_1}{C_2}\right) \frac{z}{z - 1} \tag{14.26}$$

Key Point: *Both delayed and delay-free integrators can be realized with the same switched-capacitor circuit by changing the clock phases applied to two of the switches. In a fully-differential implementation, both can be configured to provide either positive or negative dc gain.*

Finally, note that the preceding delay-free integrator has a negative gain, while the delaying integrator has a positive gain. While these sign changes in the gain can be quite useful in single-ended switched-capacitor designs, most modern integrated circuits make use of fully differential signals for improved noise performance, and sign changes can be achieved by simply cross-coupling output wires. Hence, either of the circuits in Figs. 14.9 or 14.12, when implemented fully differentially, may be used to realize either a positive or negative gain. It is therefore more specific to refer to Fig. 14.9 as a delayed integrator and Fig. 14.12 as a delay-free integrator.

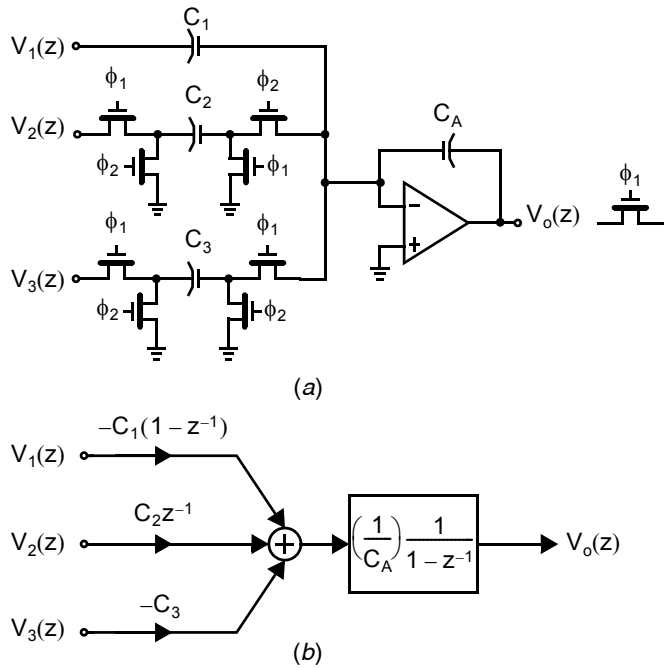


Fig. 14.14 Three-input switched capacitor summing/integrator stage. (*a*) Circuit. (*b*) Equivalent signal-flow-graph.

14.2.4 Signal-Flow-Graph Analysis

Clearly, applying the charge transfer equations just described to larger circuits would be tedious. Instead, a few rules can be developed to create and analyze such circuits graphically. Consider the three-input integrator shown in Fig. 14.14. Using the principle of superposition, it is clear that the input–output relationships for $V_2(z)$ and $V_3(z)$ are given by

$$\frac{V_o(z)}{V_2(z)} = \left(\frac{C_2}{C_A}\right)\left(\frac{z^{-1}}{1-z^{-1}}\right) \tag{14.27}$$

$$\frac{V_o(z)}{V_3(z)} = -\left(\frac{C_3}{C_A}\right)\left(\frac{1}{1-z^{-1}}\right) \tag{14.28}$$

For the input $V_1(z)$, we note that its relationship to $V_o(z)$ is simply an inverting gain stage given by

$$\frac{V_o(z)}{V_1(z)} = -\left(\frac{C_1}{C_A}\right) \tag{14.29}$$

with the output once again being sampled at the end of ϕ_1 as shown.

Combining (14.27), (14.28), and (14.29), the input–output relationship for the circuit of Fig. 14.14(*a*) is equal to

$$V_o(z) = -\left(\frac{C_1}{C_A}\right)V_1(z) + \left(\frac{C_2}{C_A}\right)\left(\frac{z^{-1}}{1-z^{-1}}\right)V_2(z) - \left(\frac{C_3}{C_A}\right)\left(\frac{1}{1-z^{-1}}\right)V_3(z) \tag{14.30}$$

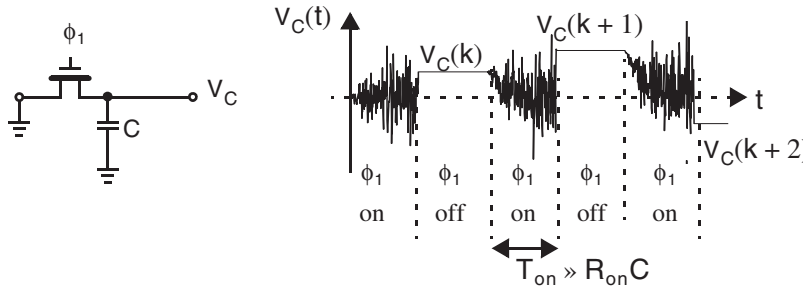


Fig. 14.15 Noise sampling on a switched capacitor.

Key Point: *Complicated switched-capacitor circuits may be analyzed graphically using a signal-flow-graph representing the opamp stage with an integrating block and replacing each input branch with the corresponding non switched, delayed, or delay-free gain factor.*

Such a relationship can also be represented by the signal-flow-graph shown in Fig. 14.14(*b*), where the input stages have been separated from the opamp stage. Thus, we see that to obtain an equivalent signal-flow-graph, the block $(1/C_A)[1/(1-z^{-1})]$ is used to represent the opamp stage, while three different gain factors are used to represent the possible input stages. For a non-switched capacitor input stage, a gain factor of $-C_1(1-z^{-1})$ is used. If a delaying switched capacitor is used, it is represented by C_2z^{-1}. Finally, for a delay-free switched-capacitor input, a gain factor of $-C_3$ is used. Note, however, that the output, $V_o(z)$, should be sampled on ϕ_1 when the switched-capacitor inputs sample their input signals on ϕ_1.

As before, this transfer function might often be rewritten in powers of z rather than z^{-1}, resulting in

$$V_o(z) = -\left(\frac{C_1}{C_A}\right)V_1(z) + \left(\frac{C_2}{C_A}\right)\left(\frac{1}{z-1}\right)V_2(z) - \left(\frac{C_3}{C_A}\right)\left(\frac{z}{z-1}\right)V_3(z) \tag{14.31}$$

Finally, it should be noted that in a fully differential implementation, effective negative capacitances for C_1, C_2, and C_3 can be achieved by simply interchanging the input wires.

14.3 NOISE IN SWITCHED-CAPACITOR CIRCUITS

The analysis of noise in switched-capacitor circuits is subtly different than that introduced earlier in this text. The noise analysis in Chapter 9 presumes continuous-time circuits under constant bias conditions, so that the random noise statistics may be considered stationary. By contrast, switched-capacitor circuits are constantly changing their configuration due to the opening and closing of switches. We will perform a simplified analysis, useful for design, by considering the circuit in each of its clock phases. Moreover, we assume that the noise voltage across capacitors is sampled instantaneously and perfectly when the switches turn off.

Key Point: *Switched capacitors accumulate the thermal noise of the switches. Sampled kT/C noise from both clock phases superimpose resulting in a total mean squared input-referred noise of 2kT/C.*

The sampling of a noise signal on a capacitor is illustrated in Fig. 14.15. After the switch turns on, thermal noise from the switch accumulates on the capacitor. For any practical switched capacitor circuit, the time constant of the switch on resistance, R_{on}, and capacitance C should be significantly less than the switch on time to ensure settling of signal voltages on the capacitor. Hence, by the time the switch opens the rms noise voltage has reached its steady state value of $\sqrt{kT/C}$ (see Chapter 9). The voltage sampled and stored on the capacitor, $V_C(k)$, is therefore a discrete-time noise voltage with an rms value of $\sqrt{kT/C}$. When an input signal is present on the left terminal of the switched capacitor in Fig. 14.15, the sampled noise voltage is superimposed on the sampled input voltage.

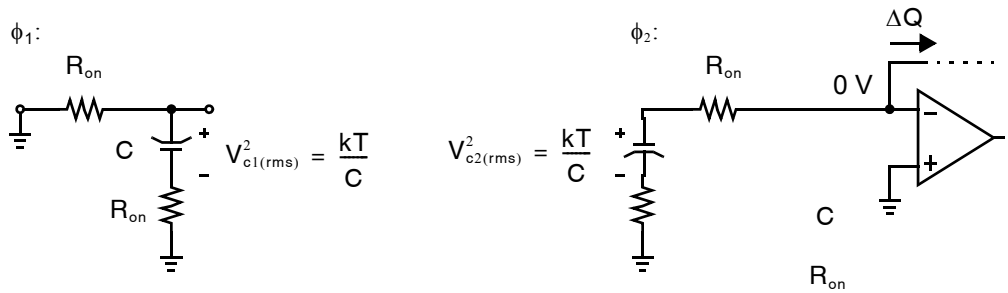


Fig. 14.16 A switched-capacitor branch shown in each clock phase with the switch resistances .R_{on}

Next, consider the delayed parasitic-insensitive switched capacitor introduced in Fig. 14.9. In Fig. 14.16, the switched-capacitor branch is illustrated in each of its two clock phases with the switches replaced by their on resistance R_{on}. Since we are presently interested only in the noise, we assume the input voltage to the branch is zero (ground). Noise is contributed by each of the four switches. In clock phase ϕ_1, a rms voltage $V^2_{c1(rms)} = \sqrt{kT/C}$ is sampled upon the capacitor. The charge stored on the capacitor is, therefore, a random quantity with a rms value

$$\Delta Q_{1(rms)} = C \cdot \sqrt{kT/C} = \sqrt{kTC} \tag{14.32}$$

This random quantity of charge is completely delivered to the opamp feedback network during clock phase ϕ_2.

Superimposed upon this is another random quantity of charge due to the switches on during clock phase ϕ_2. Again, thermal noise in the switches gives rise to a capacitor noise voltage $V^2_{c2.(rms)} = \sqrt{kT/C}$. Assuming the opamp is ideal, and therefore able to maintain a precise virtual ground during ϕ_2, this capacitor noise voltage results in an additional random charge being delivered to (or from) the feedback network.

$$\Delta Q_{2(rms)} = C \cdot \sqrt{kT/C} = \sqrt{kTC} \tag{14.33}$$

The noise quantities V_{c1} and V_{c2} are uncorrelated, as are the charges ΔQ_1 and ΔQ_2. Therefore, the mean square total charge delivered by the switched capacitor branch is a sum of squares of (14.32) and (14.33).

$$\Delta Q_{(rms)}{}^2 = \Delta Q_{1(rms)}{}^2 + \Delta Q_{2(rms)}{}^2 = 2kTC \tag{14.34}$$

This is precisely the same charge that would be delivered if the switches were noiseless and the input mean squared voltage were $2kT/C$. Hence, assuming an ideal opamp, the switched capacitor branch may be considered to have an input-referred noise,

$$V_{in(rms)} = \sqrt{\frac{2kT}{C}} \tag{14.35}$$

It may be shown that including nonideal effects of the opamp makes the input-referred noise slightly higher than the value in (14.35) but the difference is generally small when the opamp's noise is dominated by its input differential pair transistors [Schreier, 2005].

14.4 FIRST-ORDER FILTERS

Key Point: *Beginning with conventional continuous-time active RC filters, and replacing each resistor with a switched-capacitor results in a discrete-time filter whose frequency response closely approximates that of the original continuous-time filter at frequencies far below the sampling frequency. Once a filter structure is obtained, its precise frequency response is determined through the use of discrete-time analysis.*

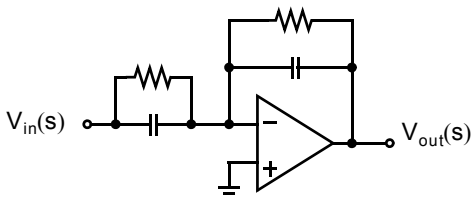
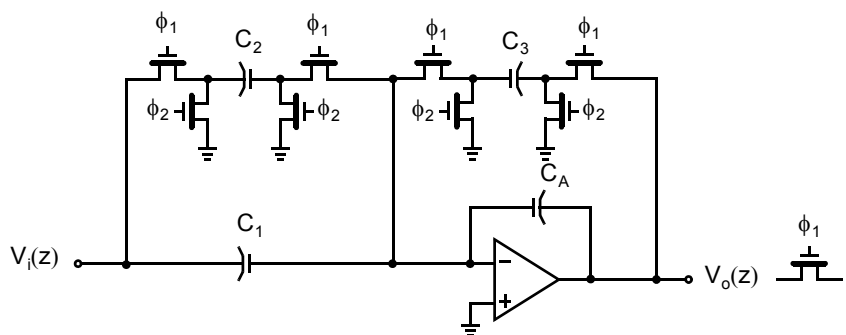


Fig. 14.17 A first-order active-RC filter.

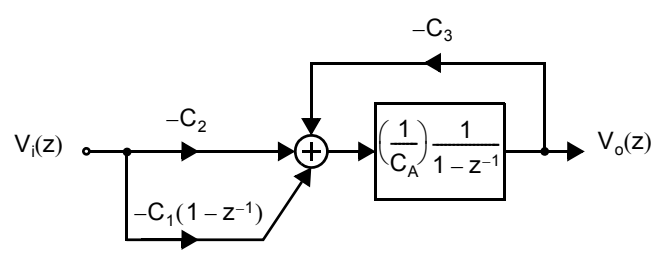
We saw in Section 14.2.1 that at low frequencies, a switched capacitor is equivalent to a resistor. Such an equivalence is often used to derive good switched-capacitor filter structures from their active-RC counterparts. However, although such structures would behave very similarly for low-frequency input signals (when compared to the clock frequency), for signal frequencies moderately close to the clock frequency, the behavior of the switched-capacitor circuit is precisely determined through the use of the z-domain signal-flow-graph approach just discussed.

A general first-order active-RC filter is shown in Fig. 14.17. To obtain a switched-capacitor filter having the same low-frequency behavior, the resistors are replaced with delay-free switched capacitors (delay-free inputs are used since they create negative integration, as do resistive inputs), while the nonswitched capacitor feed-in is left unchanged. The resulting first-order switched-capacitor filter and its equivalent signal-flow-graph are shown in Fig. 14.18. An equation that describes this signal-flow-graph is

$$C_A(1 - z^{-1})V_o(z) = -C_3V_o(z) - C_2V_i(z) - C_1(1 - z^{-1})V_i(z) \tag{14.36}$$



(*a*)



(*b*)

Fig. 14.18 A first-order switched-capacitor filter: (*a*) circuit (no switch sharing), (*b*) equivalent signal-flow-graph.

from which the transfer function for this first-order filter is found to be

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\frac{\left(\frac{C_1}{C_A}\right)(1 - z^{-1}) + \left(\frac{C_2}{C_A}\right)}{1 - z^{-1} + \frac{C_3}{C_A}} = -\frac{\left(\frac{C_1 + C_2}{C_A}\right)z - \frac{C_1}{C_A}}{\left(1 + \frac{C_3}{C_A}\right)z - 1} \tag{14.37}$$

The pole of (14.37) is found by equating the denominator to zero, which results in a pole location, z_p, given by

$$z_p = \frac{C_A}{C_A + C_3} \tag{14.38}$$

Note that for positive capacitance values, this pole is restricted to the real axis between zero and one, and therefore the circuit is always stable. In fact, with $C_3 = 0$ the circuit's pole is at $z = 1$ or, in other words, it becomes a discrete-time integrator, as expected. The zero of (14.37) is found by equating the numerator to zero, which results in a zero location, z_z, given by

$$z_z = \frac{C_1}{C_1 + C_2} \tag{14.39}$$

Therefore, for positive capacitance values, this zero location is also restricted to the real axis between zero and one. Additionally, the dc gain for this circuit is found by setting $z = 1$, which results in the dc gain being given by

$$H(1) = \frac{-C_2}{C_3} \tag{14.40}$$

Recall that in a fully differential implementation, effective negative capacitances can be realized for C_1, C_2, and C_3 by interchanging the input wires. For example, a zero at $z = -1$ could be realized by setting

$$C_1 = -0.5C_2 \tag{14.41}$$

In other words, the pair of capacitors representing C_1 would be half the size of those representing C_2 and the input wires interchanged into the C_1 pair such that $-V_i(z)$ is applied to that input branch.

EXAMPLE 14.3

Find the capacitance values needed for a first-order switched-capacitor circuit such that its 3-dB point is at 10 kHz when a clock frequency of 100 kHz is used. It is also desired that the filter have zero gain at 50 kHz and the dc gain be unity. Assume $C_A = 10$ pF.

Solution

First note that having a zero gain at 50 kHz is equivalent to placing a zero at -1. Now, making use of the bilinear transform $p = (z - 1)/(z + 1)$ as described in Section 13.5.4, the zero at -1 is mapped to $\Omega = \infty$. In addition, the frequency warping between the two domains, $\Omega = \tan(\omega/2)$, maps the –3-dB frequency of 10 kHz (or 0.2π rad/sample) in the z-domain to

$$\Omega = \tan\left(\frac{0.2\pi}{2}\right) = 0.3249$$

in the continuous-time domain. Therefore, our specification is now simplified to find the pole location for a continuous-time filter having a –3-dB frequency equal to 0.3249 rad/s and a zero at ∞. Clearly, the continuous-time pole, p_p, required here is

$$p_p = -0.3249$$

This pole is mapped back to a z-domain pole, z_p, of value

$$z_p = \frac{1 + p_p}{1 - p_p} = 0.5095$$

Therefore, the desired discrete-time transfer function, $H(z)$, is given by

$$H(z) = \frac{k(z + 1)}{z - 0.5095}$$

where k is determined by setting the dc gain to one (i.e., $H(1) = 1$) resulting in

$$H(z) = \frac{0.24525(z + 1)}{z - 0.5095}$$

or equivalently,

$$H(z) = \frac{0.4814z + 0.4814}{1.9627z - 1}$$

Equating these coefficients with those of (14.37) (and assuming $C_A = 10$ pF) results in

$$C_1 = 4.814 \text{ pF}$$
$$C_2 = -9.628 \text{ pF}$$
$$C_3 = 9.628 \text{ pF}$$

where a differential input can realize the negative capacitance.

For designs where the zero and pole frequencies are substantially less than the clock frequency (unlike the previous example, which had a zero at one-half the clock frequency), it is not necessary to pre-warp the specifications using the bilinear transform. Rather, some approximations can be used to derive equations for the capacitor ratios given the desired zero and pole, which are both assumed to be real and positive. As in Example 14.2, using (14.37), we have

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\frac{\left(\frac{C_1}{C_A}\right)(z^{1/2} - z^{-1/2}) + \left(\frac{C_2}{C_A}\right)z^{1/2}}{z^{1/2} - z^{-1/2} + \frac{C_3}{C_A}z^{1/2}} \tag{14.42}$$

Substituting (14.15) and (14.16) into (14.42) gives

$$H(e^{j\omega T}) \equiv \frac{V_o(e^{j\omega T})}{V_i(e^{j\omega T})} = -\frac{j\frac{2C_1 + C_2}{C_A}\sin\left(\frac{\omega T}{2}\right) + \frac{C_2}{C_A}\cos\left(\frac{\omega T}{2}\right)}{j\left(2 + \frac{C_3}{C_A}\right)\sin\left(\frac{\omega T}{2}\right) + \frac{C_3}{C_A}\cos\left(\frac{\omega T}{2}\right)} \tag{14.43}$$

This transfer function is exact. Making the approximation $\omega T << 1$ allows (14.43) to be simplified as

$$H(e^{j\omega T}) \equiv \frac{V_o(e^{j\omega T})}{V_i(e^{j\omega T})} = -\frac{j\left(\frac{C_1 + C_2/2}{C_A}\right)\omega T + \frac{C_2}{C_A}}{j\left(1 + \frac{C_3}{2C_A}\right)\omega T + \frac{C_3}{C_A}} \tag{14.44}$$

The procedure just used is quite useful when finding a continuous-time transfer function that approximates a discrete-time transfer function. Setting the numerator of (14.44) to zero gives the following approximate equation for the zero frequency:

$$j\omega_z T = \frac{-C_2/C_1}{1 + \frac{C_2}{2C_1}} \tag{14.45}$$

Similarly, setting the denominator to zero gives the following approximate equation for the pole frequency:

$$j\omega_p T = \frac{-C_3/C_A}{1 + \frac{C_3}{2C_A}} \tag{14.46}$$

EXAMPLE 14.4

Find the capacitance values C_3 needed for a first-order low-pass switched-capacitor filter that has $C_1 = 0$ and a pole at the 1/64th sampling frequency using the approximate equations. The low-frequency gain should be unity.

Solution

Using (14.46) and solving for C_3/C_A, we have

$$\frac{C_3}{C_A} = \frac{\omega_p T}{1 - \omega_p T/2} = \frac{2\pi/64}{1 - 2\pi/128} = 0.1032 \tag{14.47}$$

If we choose $C_A = 10$ pF, then we have $C_3 = 1.032$ pF. From (14.44) with $C_1 = 0$, we see that the low-frequency gain is given by C_2/C_3. For unity gain, this means we need $C_2 = 1.032$ pF as well.

14.4.1 Switch Sharing

Careful examination of the switched-capacitor filter shown in Fig. 14.18(*a*) reveals that some of the switches are redundant. Specifically, the top plate of both C_2 and C_3 are always switched to the opamp's virtual ground and true ground at the same time. Therefore, one pair of these switches can be eliminated and the two top plates of C_2 and C_3 can be connected together, as shown in Fig. 14.19.

14.4.2 Fully Differential Filters

While the circuits seen so far have all been shown with single-ended signals, in most analog applications it is desirable to keep the signals fully differential. Fully differential signals imply that the *difference between two lines represents the signal component,* and thus any noise which appears as a common-mode signal on those two lines does

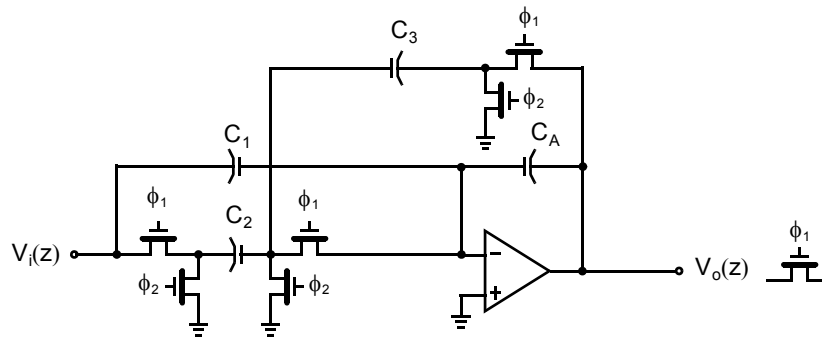


Fig. 14.19 A first-order switched-capacitor filter with switch sharing.

not affect the signal. Fully differential circuits should also be balanced, implying that the differential signals operate symmetrically around a dc common-mode voltage (typically, analog ground or halfway between the power-supply voltages). Having balanced circuits implies that it is more likely noise will occur as a common-mode signal due to the symmetry of the circuit. Fully differential circuits have the additional benefit that if each single-ended signal is distorted symmetrically around the common-mode voltage, the differential signal will have only odd-order distortion terms (which are often much smaller). To see this distortion improvement, consider the block diagram shown in Fig. 14.20, where the two nonlinear elements are identical and, for simplicity, the common-mode voltage is assumed to be zero. If the nonlinearity is memoryless, then the outputs can be found as a Taylor series expansion given by

$$v_o = k_1 v_i + k_2 v_i^2 + k_3 v_i^3 + k_4 v_i^4 + \cdots \tag{14.48}$$

Key Point: *Fully-differential filter implementations offer several advantages over their single-ended counterparts: they allow for the realization of sign inversion (i.e. a gain of –1) by simply interchanging wires, reject common-mode noise, and rejecting even-order distortion terms which generally dominate single-ended circuits.*

where k_i are constant terms. In this case, the differential output signal, v_{diff}, consists of only the linear term, $2k_1v_1$, and odd-order distortion terms (which are typically smaller than the second-order term). With these two important advantages, most modern switched-capacitor circuits are realized using fully differential structures.

A fully differential realization of the first-order filter is shown in Fig. 14.21. Note here that the fully differential version is essentially two copies of the single-ended version, which might lead one to believe that it would consume twice the amount of integrated area. Fortunately, the increased area penalty is not that high. First, we see that only one opamp is needed although it does require extra common-mode feedback circuitry. Second, note that the input and output signal swings have now been doubled in size. For example, if single-ended

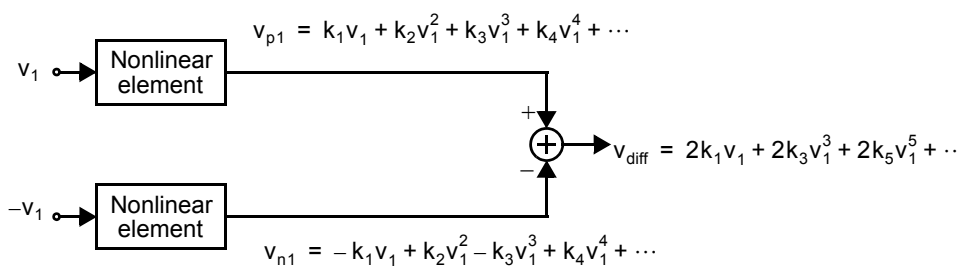

$v_{p1} = k_1v_1 + k_2v_1^2 + k_3v_1^3 + k_4v_1^4 + \cdots$
v_1
Nonlinear element
+
+
$v_{diff} = 2k_1v_1 + 2k_3v_1^3 + 2k_5v_1^5 + \cdots$
–
$-v_1$
Nonlinear element
$v_{n1} = -k_1v_1 + k_2v_1^2 - k_3v_1^3 + k_4v_1^4 + \cdots$

Fig. 14.20 Demonstrating that even-order distortion terms cancel in fully differential circuits if the distortion is symmetrical around the common-mode voltage. Here, the common-mode voltage is assumed to be zero.

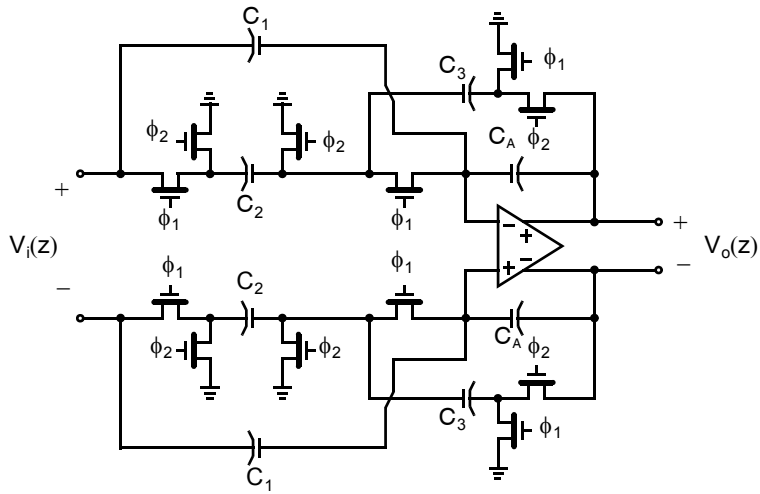


Fig. 14.21 A fully differential first-order switched-capacitor filter.

signals are restricted to ± 1 V, then the maximum peak-to-peak signal swing of single-ended signals is 2 V, while for differential signals it is 4 V. Thus, to maintain the same dynamic range due to $(kT)/C$ noise, the capacitors in the fully differential version can be half the size of those in the single-ended case.[3] Since smaller capacitors can be used, the switch size widths may also be reduced to meet the same settling-time requirement. However, the fully differential circuit has more switches and wiring so this circuit will be somewhat larger than its single-ended counterpart, although if properly designed it should be significantly less than two times larger. Moreover, recall that the fully differential circuit has the advantages of rejecting much more common-mode noise and better distortion performance.

Key Point: *Although fully-differential circuit implementations require some overhead for additional wiring and common-mode feedback circuitry, if properly designed the overhead should represent a small fraction of the overall circuit.*

14.5 BIQUAD FILTERS

Similar to the first-order case, good switched-capacitor biquad filter structures can be obtained by emulating well-known continuous-time filter structures. However, also as in the first-order case, once a filter structure is obtained, its precise frequency response is determined through the use of discrete-time analysis using the signal-flow-graph technique discussed in Section 14.2.4. This exact transfer function, or a close approximation to it, is then used when determining capacitor ratios during the design phase.

14.5.1 Low-Q Biquad Filter

A direct-form continuous-time biquad structure can be obtained by rewriting the general biquad transfer function, $H_a(s)$, as follows:

3. When the signal voltage doubles, the signal power goes up by a factor of four. Fully differential circuits would have twice the noise power of single-ended circuits having the same size capacitors due to there being two capacitive networks generating noise. Halving the capacitor size of the differential circuit would result in another doubling in noise power, resulting in a signal-to-noise ratio equal to that of a single-ended circuit having capacitors twice as large.

$$H_a(s) \equiv \frac{V_{out}(s)}{V_{in}(s)} = -\frac{k_2 s^2 + k_1 s + k_0}{s^2 + \left(\frac{\omega_0}{Q}\right)s + \omega_0^2} \tag{14.49}$$

Here, ω_0 and Q are the pole frequency and pole Q, respectively, whereas k_0, k_1, and k_2 are arbitrary coefficients that place this biquad's zeros (i.e., the numerator's roots). Multiplying through by the denominators and dividing by s^2, (14.49) can be written as

$$V_{out}(s) = \left[-k_2 - \frac{k_1}{s}\right]V_{in}(s) - \frac{\omega_0}{Qs}V_{out}(s) - \frac{k_0}{s^2}V_{in}(s) - \frac{\omega_0^2}{s^2}V_{out}(s) \tag{14.50}$$

Finally, (14.50) can be expressed as two integrator-based equations. Specifically,

$$V_{c1}(s) \equiv -\frac{1}{s}\left[\frac{k_0}{\omega_0}V_{in}(s) + \omega_0 V_{out}(s)\right] \tag{14.51}$$

and

$$V_{out}(s) = -\frac{1}{s}\left[(k_1 + k_2 s)V_{in}(s) + \frac{\omega_0}{Q}V_{out}(s) - \omega_0 V_{c1}(s)\right] \tag{14.52}$$

A signal-flow-graph describing the preceding two equations is shown in Fig. 14.22.

Now, allowing negative resistors to be used, an active-RC realization of this general biquad signal-flow-graph is shown in Fig. 14.23 [Brackett, 1978]. Note that for convenience the integrating capacitors, C_A and C_B, have been set to unity.

A switched-capacitor biquad based on this active-RC circuit can be realized as shown in Fig. 14.24 [Martin, 1979]. Here, all positive resistors were replaced with delay-free feed-in switched-capacitor stages while the negative resistor was replaced with a delaying feed-in stage. The switched-capacitor circuit has redundant switches, as switch sharing has not yet been applied.

The input capacitor K_1C_1 is the major signal path when realizing low-pass filters; the input capacitor K_2C_2 is the major signal path when realizing bandpass filters; whereas, the input capacitor K_3C_2 is the major signal path when realizing high-pass filters.

Before proceeding, it is worth mentioning that while this switched-capacitor filter is a natural result when using single-ended circuits, the use of fully differential circuits could result in quite a few other circuits, some of which could have poor performance properties. For example, all the positive resistors could be replaced with delay-free feed-in stages as before, but in addition, the negative resistor could also be replaced with a delay-free

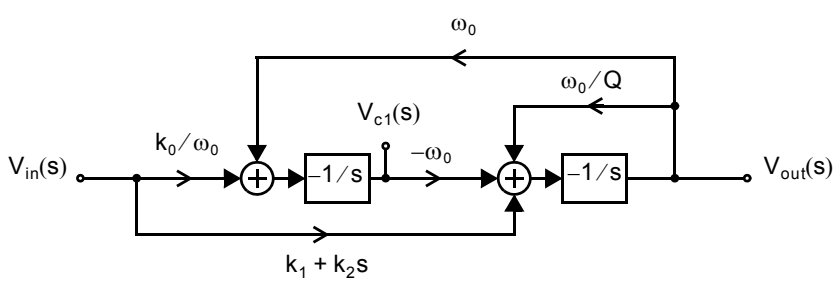


Fig. 14.22 A signal-flow-graph representation of a general continuous-time biquad filter.

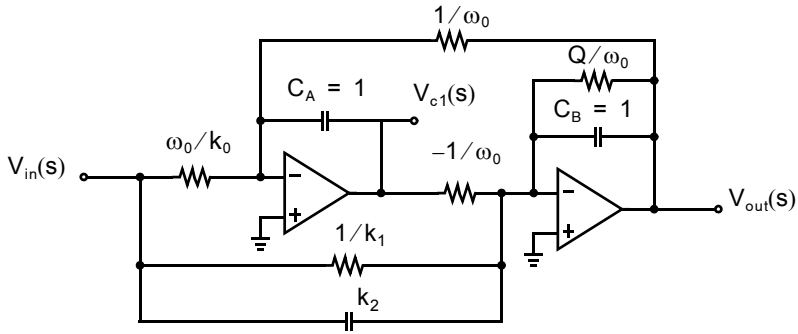


Fig. 14.23 An active-RC realization of a general continuous-time biquad filter.

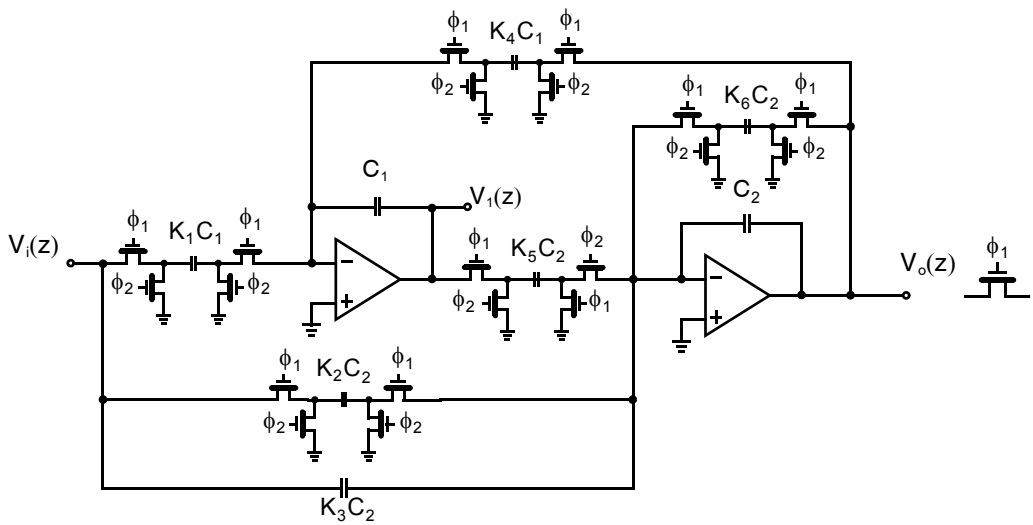


Fig. 14.24 A low-Q switched-capacitor biquad filter (without switch sharing).

feed-in stage where the differential input wires are interchanged. Such a circuit would have a delay-free loop around the two integrators and may have an excessive settling time behavior. As another example, all resistors might be replaced with delaying feed-in stages where positive resistors are realized by interchanging the input wires of those stages. In this case, settling time behavior would not suffer as there would be two delays around the two-integrator loop, but coefficient sensitivity would be worse for filters having high-Q poles. In summary, what is important is that the two-integrator loop have a single delay around the loop. Such an arrangement is referred to as using lossless discrete integrators (LDI) [Bruton, 1975].

Using the approach presented in Section 14.2.4, the signal-flow-graph for the biquad circuit shown in Fig. 14.24 can be found and is shown in Fig. 14.25. The transfer function, $H(z)$, for this signal-flow-graph is found to be given by

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\frac{(K_2 + K_3)z^2 + (K_1K_5 - K_2 - 2K_3)z + K_3}{(1 + K_6)z^2 + (K_4K_5 - K_6 - 2)z + 1} \tag{14.53}$$

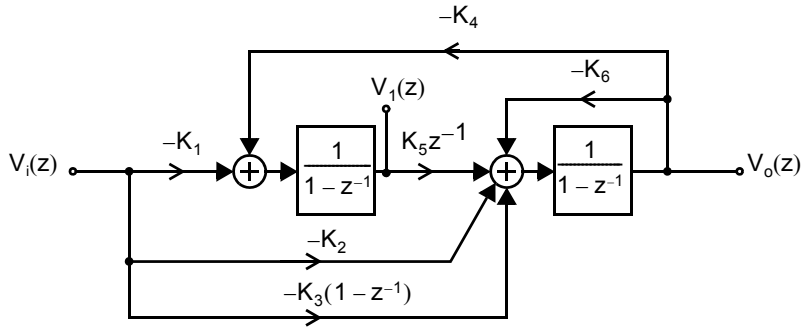


Fig. 14.25 A signal-flow-graph describing the switched-capacitor circuit in Fig. 14.24.

While the preceding relation is useful for analyzing the switched-capacitor circuit shown, this relation can also be used during design of a desired transfer function. Assuming the desired discrete-time biquad transfer function has been found using a digital-filter design program, and is given by

$$H(z) = -\frac{a_2z^2 + a_1z + a_0}{b_2z^2 + b_1z + 1} \tag{14.54}$$

we can equate the individual coefficients of z in (14.53) and (14.54), resulting in the following design equations:

$$K_3 = a_0 \tag{14.55}$$

$$K_2 = a_2 - a_0 \tag{14.56}$$

$$K_1K_5 = a_0 + a_1 + a_2 \tag{14.57}$$

$$K_6 = b_2 - 1 \tag{14.58}$$

$$K_4K_5 = b_1 + b_2 + 1 \tag{14.59}$$

There is some flexibility in choosing the values of K_1, K_4, and K_5. A single degree of freedom arises because specifying the overall transfer function, $H(z)$, places no constraint on the size of the signal at the internal node, $V_1(z)$. Thus, one way to choose appropriate capacitor values is to initially let $K_5 = 1$ and determine other initial component values, find the signal levels that occur at nodes $V_1(z)$ and $V_o(z)$, then using that signal level information, perform dynamic range scaling to find the final component values that result in similar signal swings at both nodes. Alternatively, the time constants of the two discrete-time integrators can be set equal by choosing

$$K_4 = K_5 = \sqrt{b_1 + b_2 + 1} \tag{14.60}$$

Such a choice usually results in a near optimal dynamically range-scaled circuit.

If it is desired to design the SC biquad so that its transfer function matches that of a continuous-time transfer function as closely as possible, then the approach used in Example 14.2 can be taken. Specifically, (14.53) can be rewritten as

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\frac{K_1K_5 + K_2(z^{1/2} - z^{-1/2})z^{1/2} + K_3(z^{1/2} - z^{-1/2})^2}{K_4K_5 + K_6(z^{1/2} - z^{-1/2})z^{1/2} + (z^{1/2} - z^{-1/2})^2} \tag{14.61}$$

Rewriting (14.15) and (14.16) here for convenience, we have

$$z^{1/2} = \cos\left(\frac{\omega T}{2}\right) + j\sin\left(\frac{\omega T}{2}\right) \tag{14.62}$$

and

$$z^{-1/2} = \cos\left(\frac{\omega T}{2}\right) - j\sin\left(\frac{\omega T}{2}\right) \tag{14.63}$$

Substituting these equations into (14.61) and rearranging gives

$$H(e^{j\omega T}) \equiv \frac{V_o(e^{j\omega T})}{V_i(e^{j\omega T})} = -\frac{K_1K_5 + jK_2\ \sin(\omega T) + (4K_3 + 2K_2)\sin^2\left(\frac{\omega T}{2}\right)}{K_4K_5 + jK_6\ \sin(\omega T) + (4 + 2K_6)\sin^2\left(\frac{\omega T}{2}\right)} \tag{14.64}$$

For $\omega T << 1$, we further have

$$H(e^{j\omega T}) \equiv \frac{V_o(e^{j\omega T})}{V_i(e^{j\omega T})} \approx -\frac{K_1K_5 + jK_2(\omega T) + (K_3 + K_2/2)(\omega T)^2}{K_4K_5 + jK_6(\omega T) + (1 + K_6/2)(\omega T)^2} \tag{14.65}$$

The coefficients of (14.65) can be matched to the desired coefficients of the continuous-time transfer function. A closer matching between transfer functions can be obtained if one chooses capacitor ratios to match the real and imaginary parts of both the numerators and denominators of (14.64) and the desired continuous-time transfer function at the resonant frequency, as is outlined in [Martin, 1980].

Finally, an estimate of the largest to smallest capacitance ratio can be made by using the resistor approximation (14.5) for this switched-capacitor circuit. Comparing Fig. 14.23 and Fig. 14.24 and using (14.5), we see the following approximations are valid for the feedback capacitor ratios, K_4, K_5, and K_6 (the capacitor ratios that determine the pole locations).

$$K_4 \cong K_5 \cong \omega_0 T \tag{14.66}$$

$$K_6 \cong \frac{\omega_0 T}{Q} \tag{14.67}$$

However, the sampling rate, $1/T$, is typically much larger that the approximated pole frequency, ω_0, resulting in

$$\omega_0 T << 1 \tag{14.68}$$

Thus, the largest capacitors (that determine the pole locations) in the switched-capacitor biquad circuit of Fig. 14.24 are the integrating capacitors, C_1 and C_2. Additionally, if $Q < 1$, then the smallest capacitors are K_4C_1 and K_5C_2, resulting in an approximate capacitance spread of $1/(\omega_0 T)$. However, if $Q > 1$, then from (14.67) the smallest capacitor would be K_6C_2, resulting in an approximate capacitance spread of $Q/(\omega_0 T)$.

14.5.2 High-Q Biquad Filter

The resultant high-capacitance spread in the preceding low-Q circuit, when $Q >> 1$, was a direct result of having a large damping resistance value of Q/ω_0 in Fig. 14.23. One way to obtain a high-Q switched-capacitor circuit that does not have such a high capacitance spread is to eliminate this damping resistance in the continuous-time filter counterpart. Such an elimination can be accomplished by rewriting (14.50) as

$$V_{out}(s) = -\frac{1}{s}[k_2 s V_{in}(s) - \omega_0 V_{c1}(s)] \tag{14.69}$$

and

$$V_{c1}(s) = -\frac{1}{s}\left[\left(\frac{k_0}{\omega_0} + \frac{k_1 s}{\omega_0}\right) V_{in}(s) + \omega_0 V_{out}(s) + \frac{s}{Q} V_{out}(s)\right] \tag{14.70}$$

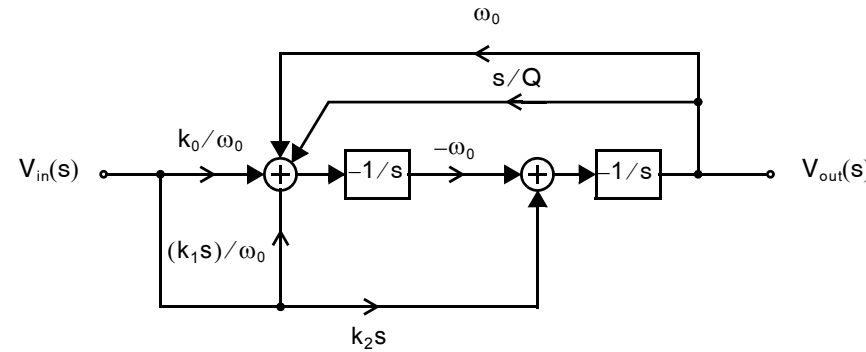


Fig. 14.26 An alternate realization for a general continuous-time biquad filter.

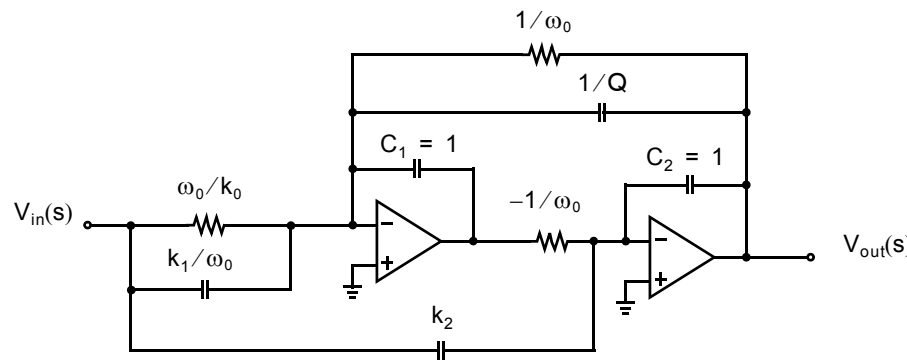


Fig. 14.27 An alternate realization of a general active-RC biquad filter.

The signal-flow-graph representation of these two equations is shown in Fig. 14.26. Note that here the damping path, s/Q coefficient, passes through the first integrator stage and can be realized through the use of a capacitor, as shown in Fig. 14.27.

Once again, replacing resistors with appropriate switched-capacitor feed-in stages, the final high-Q switched-capacitor biquad circuit is shown in Fig. 14.28 [Martin, 1980].

The input capacitor K_1C_1 is the major signal path when realizing low-pass filters, the input capacitor K_2C_1 is the major signal path when realizing bandpass filters, and the input capacitor K_3C_2 is the major signal path when realizing high-pass filters. Other possibilities for realizing different types of functions exist. For example, a nondelayed switched capacitor going from the input to the second integrator could also be used to realize an inverting bandpass function. If the phases on the input switches of such a switched capacitor were interchanged, then a noninverting bandpass function would result, but with an additional period delay.

Using the signal-flow-graph approach described in Section 14.2.4, the transfer function for this circuit is found to be given by

$$H(z) \equiv \frac{V_o(z)}{V_i(z)} = -\frac{K_3z^2 + (K_1K_5 + K_2K_5 - 2K_3)z + (K_3 - K_2K_5)}{z^2 + (K_4K_5 + K_5K_6 - 2)z + (1 - K_5K_6)} \tag{14.71}$$

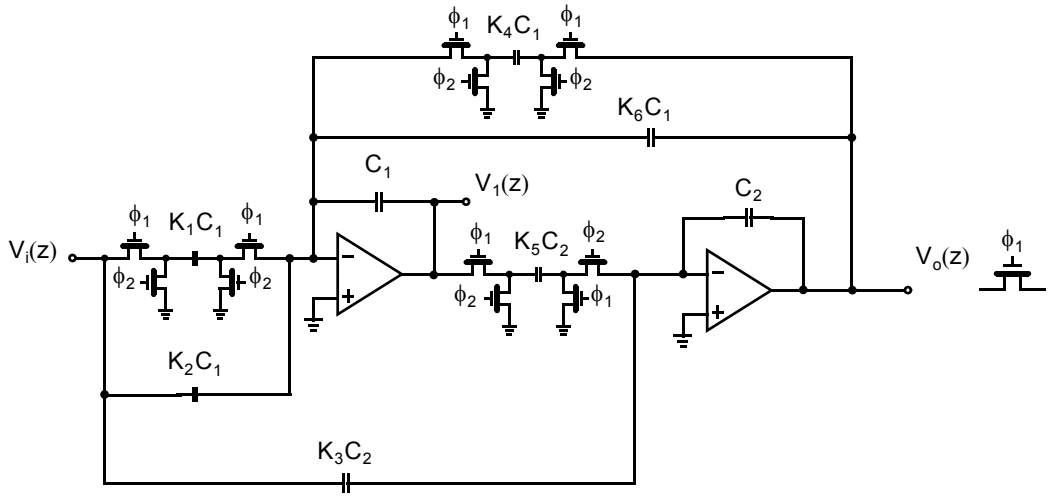


Fig. 14.28 A high-Q switched-capacitor biquad filter (without switch sharing).

Matching the coefficients of (14.71) to the coefficients of a discrete-time transfer function given by[4]

$$H(z) = -\frac{a_2z^2 + a_1z + a_0}{z^2 + b_1z + b_0} \tag{14.72}$$

the following equations result:

$$K_1K_5 = a_0 + a_1 + a_2 \tag{14.73}$$

$$K_2K_5 = a_2 - a_0 \tag{14.74}$$

$$K_3 = a_2 \tag{14.75}$$

$$K_4K_5 = 1 + b_0 + b_1 \tag{14.76}$$

$$K_5K_6 = 1 - b_0 \tag{14.77}$$

As for the low-Q biquad, we see that there is some freedom in determining the coefficients as there is one less equation than the number of coefficients. A reasonable first choice might be to take

$$K_4 = K_5 = \sqrt{1 + b_0 + b_1} \tag{14.78}$$

after which all other ratios follow. This will give a good dynamic range but not necessarily optimum. After the biquad has been designed, it can be analyzed and the ratios adjusted to equalize the maximum signal strength at the two opamp outputs.

EXAMPLE 14.5

The following transfer function describes a bandpass filter having a peak gain of 5 near $f_s/10$ and a Q of about 10.

$$H(z) = -\frac{0.288(z - 1)}{z^2 - 1.572z + 0.9429} \tag{14.79}$$

4. Note that (14.72) is written in slightly different form than (14.54) to simplify the resulting equations.

Find the largest to smallest capacitor ratio if this transfer function is realized using the high-Q biquad circuit. Compare the results to those which would be obtained if the low-Q circuit were used instead. Let $C_1 = C_2 = 1$ in both cases.

Solution

For the high-Q biquad circuit, making use of (14.73) to (14.76) and assuming $K_4 = K_5$ results in the following capacitor values.

$$K_4 = K_5 = 0.6090$$

$$K_1 = 0$$

$$K_2 = 0.4729$$

$$K_3 = 0$$

$$K_6 = 0.0938$$

Recalling that $C_1 = C_2 = 1$, in this high-Q circuit the smallest capacitor is K_6C_1, which makes the capacitance spread about 11 to 1.

For the low-Q biquad circuit design, we first rewrite (14.79) as

$$H(z) = -\frac{0.3054(z-1)\,z}{1.0606z^2 - 1.667z + 1} \tag{14.80}$$

which makes applying equations (14.55) to (14.60) easier. In addition, an extra zero at $z = 0$ was added to avoid negative capacitances. This extra zero does not change the magnitude response.

We obtain the following ratios assuming $K_4 = K_5$:

$$K_4 = K_5 = 0.6274 \tag{14.81}$$

$$K_1 = 0 \tag{14.82}$$

$$K_2 = 0.3054 \tag{14.83}$$

$$K_3 = 0 \tag{14.84}$$

$$K_6 = 0.0606 \tag{14.85}$$

Thus the capacitance ratio for this circuit is determined by K_6, which primarily sets the accuracy of the pole-Q factor, resulting in a spread of near 17. It should be noted here that worse capacitance spreads will occur for this low-Q circuit if the pole Q is higher and/or the sampling frequency is increased relative to the passband edge.

Key Point: *Several different discrete-time biquads can be realized by varying the type (i.e., delayed, delay-free) and exact connection of switched-capacitor branches within a two-integrator loop. The choice of which biquad to use depends upon the specific pole and zero locations sought.*

There are many other biquadratic stages than the two presented here. For example, a general stage capable of realizing any z-transform is described in [Laker, 1994]. However, the authors have found that for the great majority of filtering applications, the choice of the circuits of either Fig. 14.24 or Fig. 14.28 (with switch sharing) is nearly optimum. The exception to this is for phase equalizers, where often feed-ins approximating negative resistors are needed. To realize higher-order switched-capacitor filters, one can cascade a number of biquad and/or first-order stages. Alternatively, one can use lower-sensitivity techniques such as signal-flow-graph

simulations of the relationships between the state variables of passive doubly terminated LC ladder filters. The interested reader is referred to [Gregorian, 1986] for detailed descriptions on the design of high-order SC filters.

14.6 CHARGE INJECTION

As seen in Chapter 10, comparators that make use of switches suffer from charge-injection errors. Fortunately, by turning off certain switches first, the charge-injection errors of the overall circuit are due to only those early switches. This same result applies to switched-capacitor circuits, and thus charge-injection effects can be minimized by having some clock signals slightly advanced with respect to the remaining signals. Which clock signals should be advanced is most easily explained with the following example.

Consider the first-order switched-capacitor filter shown in Fig. 14.29, where switch sharing is used. Here, ϕ_{1a} and ϕ_{2a} are slightly advanced with respect to the remaining clock signals for the following reasons. First, since ϕ_{2a} is always connected to ground while ϕ_{1a} is always connected to virtual ground, when these switches are turned on they need only pass a signal near the ground node (not rail-to-rail signals as might occur at opamp outputs). Thus, these two switches can be realized using single n-channel transistors rather than full CMOS transmission gates.[5] A second, more important reason for using this clock arrangement is that the charge injections due to Q_3, Q_4 are not signal dependent, while those of Q_1, Q_6 are signal dependent. Specifically, we saw in Chapter 10 that the channel charge of an NMOS transistor in triode is given by

$$Q_{CH} = -WLC_{ox}V_{eff} = -WLC_{ox}(V_{GS} - V_t) \tag{14.86}$$

Here, we see that the charge is related to the gate-source voltage as well as to the threshold voltage. When Q_3 and Q_4 are on, $V_{GS} = V_{DD}$, and since their source remains at zero volts, their threshold voltages also remain constant. As a result, *the amount of charge injected by* Q_3, Q_4 *is the same from one clock cycle to the next and can be considered as a dc offset*. Unfortunately, the same cannot be said of Q_1 and Q_6. For example, when Q_1 is on, its channel charge is found from (14.86) to be

$$Q_{CH1} = -W_1L_1C_{ox}(V_{DD} - V_i - V_{tn}) \tag{14.87}$$

Thus, one portion of the channel charge is linearly related to V_i. However, its source voltage settles to V_i and thus the transistor's threshold voltage changes in a nonlinear relationship (due to the bulk effect, assuming its substrate

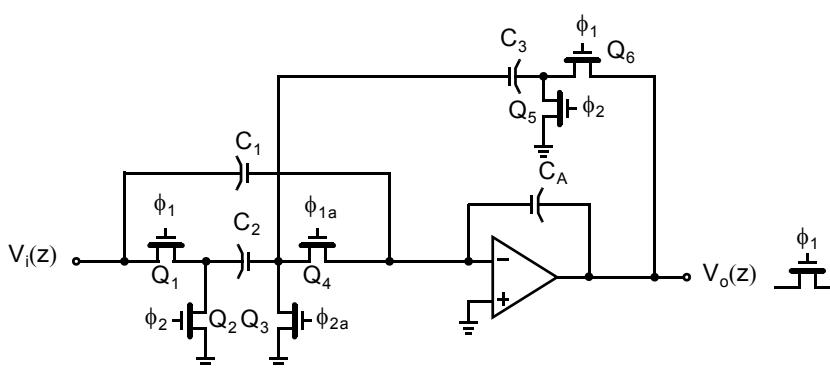


Fig. 14.29 A first-order switched-capacitor circuit with a clock arrangement to reduce charge-injection effects. ϕ_{1a} and ϕ_{2a} turn off Q_3 and Q_4 slightly early.

5. Switches Q_1, Q_6 may be realized as CMOS transmission gates since they may need to pass large-swing signals.

Key Point: *To minimize distortion due to charge injection, switches connected to dc nodes, such as ground, are turned off in advance of other switches. Although there is still some charge injection, it is signal independent.*

is set to a fixed voltage). As a result, Q_{CH1} has a linear and nonlinear relationship to V_i and thus would cause a gain error and distortion if Q_1 were turned off early. Finally, while Q_2 and Q_5 would also inject constant charge from one cycle to the next, it is best not to add any unnecessary charge into the circuit so these switches are also not advanced.

In summary, to reduce the effects of charge injection in switched-capacitor circuits, realize all switches connected to ground or virtual ground as n-channel switches only, and turn off the switches near the virtual ground node of the opamps first. Such an approach will minimize distortion and gain error as well as keeping dc offset low.

EXAMPLE 14.6

Assuming an ideal opamp, estimate the amount of dc offset in the output of the circuit shown in Fig. 14.29 due to channel-charge injection (ignore overlap capacitance charge injection) when $C_1 = 0$ and $C_2 = C_A = 10C_3 = 4$ pF. Assume that switches Q_3, Q_4 are n-channel devices with a threshold voltage of $V_{tn} = 0.45$ V, width of $W = 5$ μm, length of $L = 0.2$ μm, and $C_{ox} = 8.5$ fF/μm^2. A power supply of 1.8 V is used, and the circuit's "ground" node is located at $V_{DD}/2 = 0.9$ V.

Solution

The dc feedback around the opamp will keep the negative input of the opamp at $V_{DD}/2 = 0.9$ V. From (14.86), we can calculate the amount of channel charge of Q_3, Q_4 (when on) to be

$$Q_{CH3} = Q_{CH4} = -(5\ \mu\text{m})(0.2\ \mu\text{m})(8.5\ \text{fF}/\mu\text{m}^2)((1.8\ \text{V} - 0.9\ \text{V}) - 0.45\ \text{V}) = -3.8\ \text{fC} \quad (14.88)$$

All of the dc feedback current is charge transferred through the switched capacitor C_3. Thus, the charge being transferred through this capacitor must equal the charge injected by both Q_3, Q_4 (assuming the input is at zero volts and therefore does not contribute charge). The charge transfer into C_3 is given by

$$Q_{C_3} = -C_3 v_{out} \quad (14.89)$$

Now, we estimate that half the channel charges of Q_3 and Q_4 are injected to the virtual ground node by the following reasoning. When ϕ_{1a} turns off, half the channel charge of Q_4 goes into the virtual ground while half of its charge is placed at the node between Q_3 and Q_4. When ϕ_{2a} goes high, the second charge escapes to ground, but when ϕ_{2a} goes low, half of its channel charge is left between Q_3 and Q_4 and is passed into the virtual ground when ϕ_{1a} goes high again. As a result, we can write the following charge equation,

$$\frac{1}{2}(Q_{CH3} + Q_{CH4}) = Q_{C_3} \quad (14.90)$$

which can be used to find Q_{C_3}, and hence the output voltage from (14.89)

$$v_{out} = \frac{3.8\ \text{fC}}{0.4\ \text{pF}} = 9.5\ \text{mV} \quad (14.91)$$

Thus, we see that this output dc offset value is affected by the size of capacitors used (in particular, C_3) as well as the switch sizes and power-supply voltage. It should be mentioned here that if the dc input offset voltage of the opamp was included in the analysis, it would be multiplied by the dc gain of the circuit similar to that which occurs in an active-RC circuit.

Charge-injection is especially troublesome at higher frequencies. At higher frequencies, the time constant of the switch-on resistance and the capacitor being charged or discharged must be smaller; this necessitates larger switches for a given size capacitor and therefore greater charge-injection. It is possible to derive a very simple formula that gives an approximate upper bound on the frequency of operation of an SC circuit for a specified maximum voltage change due to charge injection. This upper bound takes into account switch channel charge only and ignores charge injection due to overlap capacitance.

Key Point: *Charge injection is most problematic at high switching speeds where low switch on resistances demand wide transistors and small capacitors.*

Most SC circuits will have two series switches for each capacitor. If one assumes that, for good settling, the sampling clock half-period must be greater than five time constants, then we have

$$\frac{T}{2} > 5R_{on}C \tag{14.92}$$

where T is the sampling period, R_{on} is the on resistance of the n-channel switch, and C is the capacitor. Equation (14.92) may be rewritten

$$f_{clk} < \frac{1}{10R_{on}C} \tag{14.93}$$

where $f_{clk} = 1/T$ is the clock frequency. Recalling (1.108) from Chapter 1, we have

$$R_{on} = \frac{1}{\mu_n C_{ox}\frac{W}{L}V_{eff}} \tag{14.94}$$

Also, using (14.87), the charge change due to the channel charge only caused by turning an n-channel switch off is approximated by

$$|\Delta V| = \frac{Q_{CH}}{2C} = \frac{WLC_{ox}V_{eff}}{2C} \tag{14.95}$$

For a specified $|\Delta V|_{max}$, (14.95) implies that

$$C = \frac{WLC_{ox}V_{eff}}{2|\Delta V|_{max}} \tag{14.96}$$

Substituting (14.94) and (14.96) into (14.93) gives

$$f_{clk} < \frac{\mu_n|\Delta V|_{max}}{5L^2} \tag{14.97}$$

a very simple expression. Thus, the upper frequency limit to charge injection for SC circuits is inversely proportional to L^2 and is approximately independent of the size of capacitors or the power-supply voltages, assuming the switch sizes are chosen optimally; as technology improves, much higher switching frequencies are attainable. Finally, (14.97) ignores overlap capacitance, so it is somewhat optimistic and should be therefore considered an upper bound on the maximum clocking frequency.

Key Point: *Technology scaling permits smaller transistors to offer the same on resistance, thereby reducing charge injection or, for the same charge injection, enabling higher switching frequencies.*

EXAMPLE 14.7

Assuming the maximum voltage change due to clock feedthrough is 1 mV, what are the maximum clocking frequencies considering only channel charge injection for technologies having minimum channel lengths of 0.8 μm, 0.35 μm, and 0.18 μm in Table 1.5?

Solution

Using (14.97) and carrier mobilities obtained from Table 1.5 gives f_{clk}, which must be less than 14.8 MHz, 69 MHz, and 196 MHz, respectively, for the three different technologies. Notice the substantial increase in speed for SC circuits as the technology improves.

14.7 SWITCHED-CAPACITOR GAIN CIRCUITS

Perhaps the most common nonfiltering analog function is a gain circuit where the output signal is a scaled version of the input. In this section, we shall see that it is possible to realize accurate gain circuits by making use of switched-capacitor techniques. An important application of switched-capacitor gain circuits is in pipelined analog-to-digital converters. Specific switched-capacitor circuits that are popular for that application are presented in the chapter on Nyquist-rate A/D converters.

14.7.1 Parallel Resistor-Capacitor Circuit

In active-RC circuits, a gain circuit can be realized as parallel combinations of resistors and capacitors in both the feedback and feed-in paths, as shown in Fig. 14.30(*a*). One of the first switched-capacitor gain circuits uses a similar approach, where the two resistors have been replaced by their switched-capacitor equivalents, as shown in Fig. 14.30(*b*) [Foxall, 1980]. Using the signal-flow-graph technique of Section 14.2.4, straightforward analysis results in the transfer function for this switched-capacitor gain circuit being given by

$$H(z) = \frac{V_{out}(z)}{V_{in}(z)} = -K \tag{14.98}$$

Unfortunately, this design will also amplify the 1/f noise and offset voltage of the shown opamp by K. However, it does have the advantage that its output is a continuous waveform that does not incur any large slew-rate requirement (as occurs in the resettable gain circuit discussed next).

14.7.2 Resettable Gain Circuit

The next configuration continuously resets an integrating capacitor on each clock cycle, as shown in Fig. 14.31 [Gregorian, 1981]. Here, the voltage across the integrating capacitor, C_2, is cleared on each ϕ_2, while on ϕ_1 the input voltage charges C_1 and the charging current flows across C_2 at the same time. In this way, the change in charge across C_2, ΔQ_{C2}, equals the change in charge across C_1, ΔQ_{C1}, and therefore, at the end of ϕ_1, the output voltage is related to the input voltage by $V_{out}/V_{in} = -C_1/C_2$.

In addition, this circuit stores any opamp input-offset voltage across the input and feedback capacitors. When the output is sampled by the succeeding circuit (during ϕ_1), the effects of the input-offset voltage of the opamp are cancelled from the output voltage. The elimination of the opamp's input-offset voltage is important for two reasons. One reason is that if the offset voltage is not cancelled, it will also be amplified when the input signal is

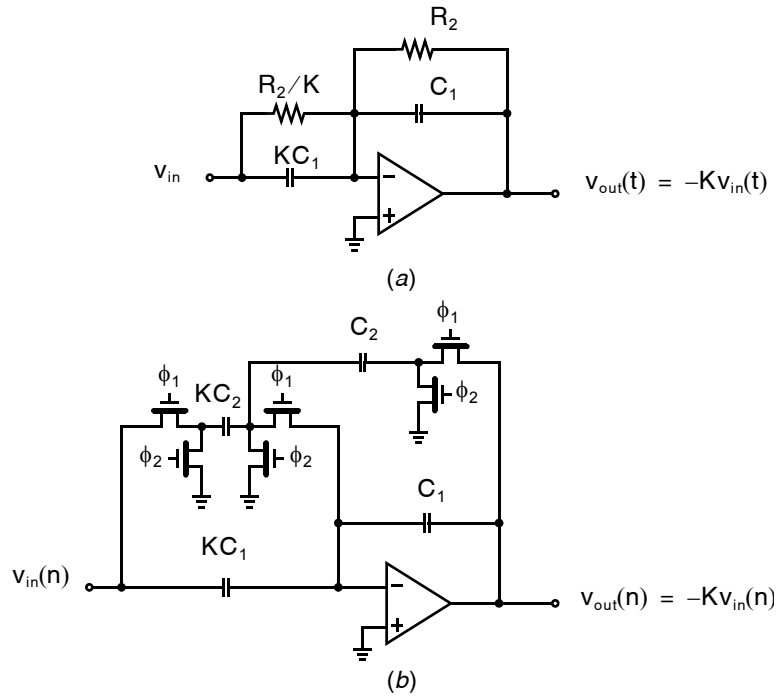


Fig. 14.30 (*a*) An active-RC gain circuit. (*b*) An equivalent switched-capacitor gain circuit.

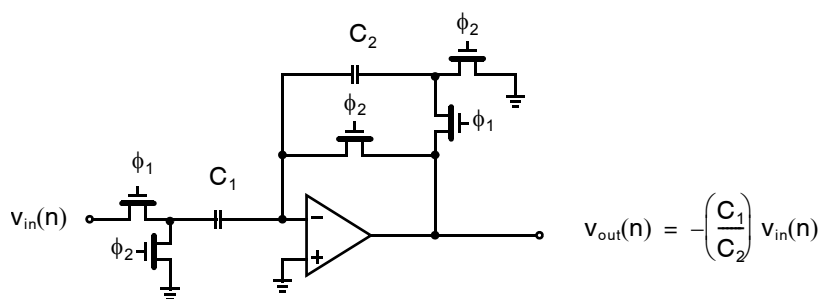


Fig. 14.31 A resettable gain circuit where opamp offset voltage is cancelled.

amplified and therefore can be troublesome. A second reason (and often more important) is that when the offset voltage is eliminated, 1/f noise is also reduced. As we saw in Chapter 9, 1/f noise is large at low frequencies and can be a dominant noise source in many MOS circuits. However, when a circuit cancels dc offset voltages, the circuit has, in effect, a high-pass response from the opamp's input terminals to the output voltage. Thus, the 1/f noise is also high-pass filtered and hence reduced. It should be noted here that although there is a high-pass response from the opamp's input terminals to the output voltage, the response from the overall circuits does not necessarily have a high-pass response.

Key Point: *Switched-capacitor circuits can use a reset phase to sample and cancel dc offset and very low-frequency (1/f) noise without affecting the circuit's overall dc gain.*

To see this offset cancellation, consider the gain circuit during ϕ_2 (i.e., when it is being reset) as shown in Fig. 14.32(*a*). The effect of the input offset voltage is being modelled as a voltage source, V_{off}, which is placed in series

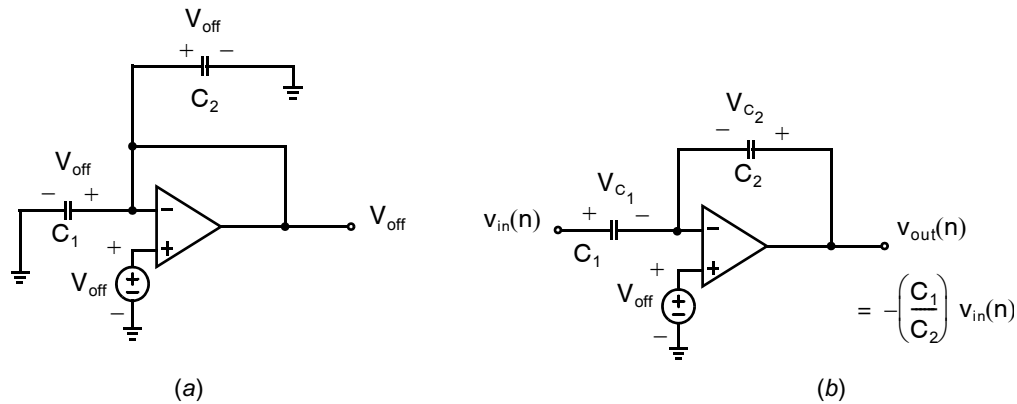


Fig. 14.32 The resettable gain circuit (*a*) during reset (ϕ_2), and (*b*) during valid output (ϕ_1).

with one of the opamp inputs. In this case, it is placed in series with the positive input, which results in the analysis being marginally simpler. During ϕ_2, both the voltages across C_1 and C_2 are equal to the opamp offset voltage, V_{off}. Next, during ϕ_1 the circuit is configured as shown in Fig. 14.32(*b*), and at the end of ϕ_1 the voltage across C_1 is given by $V_{C_1}(n) = v_{in}(n) - V_{off}$, while that for C_2 is given by $V_{C_2}(n) = v_{out}(n) - V_{off}$. Therefore we can write,

$$\Delta Q_{C1} = C_1[V_{C_1}(n) - (-V_{off})] = C_1 v_{in}(n) \tag{14.99}$$

and

$$V_{C_2}(n) = V_{C_2}(n - 1/2) - \frac{\Delta Q_{C2}}{C_2} \tag{14.100}$$

Since the voltage across C_2 one-half period earlier was $-V_{off}$ and $\Delta Q_{C2} = \Delta Q_{C1}$, we have

$$V_{C_2}(n) = -V_{off} - \frac{C_1 v_{in}(n)}{C_2} \tag{14.101}$$

Finally, since one side of C_2 is connected to the virtual ground of the opamp, which is at V_{off}, the output voltage is given by

$$\begin{aligned} v_{out}(n) &= V_{off} + V_{C_2}(n) \\ &= V_{off} + \left(-V_{off} - \frac{C_1 v_{in}(n)}{C_2}\right) \\ &= -\left(\frac{C_1}{C_2}\right) v_{in}(n) \end{aligned} \tag{14.102}$$

Thus, the output voltage is independent of the opamp offset voltage.

Note, however, that the output voltage is only valid during ϕ_1, and it is equal to V_{off} during ϕ_2, as shown in Fig. 14.33. The difficulty in realizing this waveform is that the opamp output must slew between $(-C_1/C_2)v_{in}$ and a voltage near 0 V each time the clock changes. Clearly, such a waveform requires a very high slew-rate opamp for proper operation.

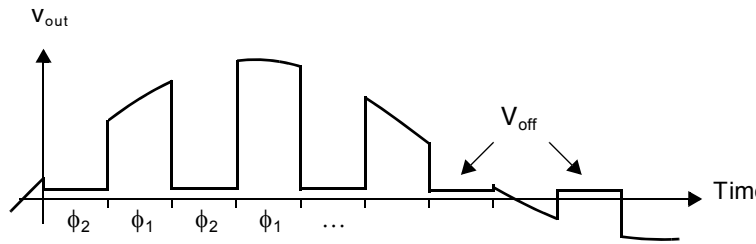


Fig. 14.33 An example output waveform for the resettable gain circuit.

Finally, note that a programmable gain circuit is easily realized by replacing C_1 with a programmable capacitor array[6] (PCA) where the capacitor size is determined by digital signals.

14.7.3 Capacitive-Reset Gain Circuit

To eliminate the need for the opamp output to slew to approximately 0 V each clock period yet still cancel the opamp's offset voltage, a capacitive-reset gain circuit can be used. The basic idea of the gain circuit is to couple the opamp's output to the inverting input during the reset phase with a capacitor that has been previously charged to the output voltage. Thus, we shall see that one property of this gain circuit is that the opamp's output need only change by the opamp's offset voltage between clock phases. In addition, since the circuit is insensitive to the opamp's input offset voltage, it also reduces the effect of the opamp's 1/f noise. Finally, it can also be shown (see [Martin, 1987]) that for low-frequency inputs (or for inputs that are constant for at least two clock periods), the errors due to finite gain, A, of the opamp are proportional to $1/A^2$ (rather than the usual $1/A$). Such a result often allows the use of single-stage opamps that are very fast opamps.

The capacitive-reset gain circuit is shown in Fig. 14.34. Capacitor C_4 is an optional "deglitching" capacitor [Matsumoto, 1987] used to provide continuous-time feedback during the nonoverlap clock times when all the

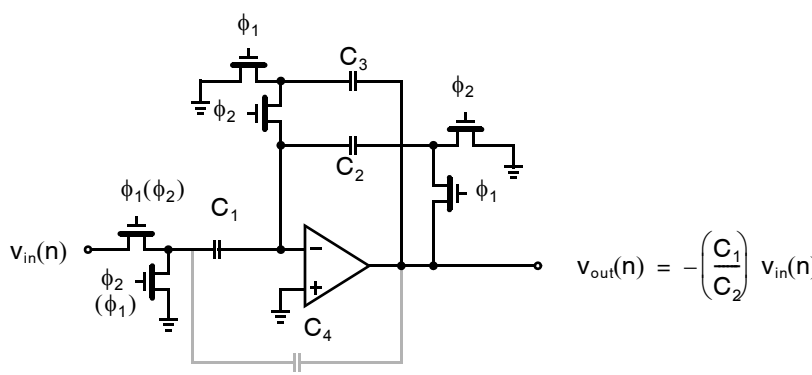


Fig. 14.34 A gain circuit using capacitive reset (C_4 is an optional deglitching capacitor). Depending on the input-stage clock signals, the gain can be either inverting (as shown) or noninverting (input-stage clocks shown in parentheses).

6. A programmable capacitor array is a capacitor whose size can be digitally controlled. It normally consists of a number of binarily-weighted capacitors having one plate common for all capacitors and the other plate of each capacitor connected through digitally-controlled switches to either a second common node or ground.

switches are open. This deglitching technique works well and should be used on almost all switched-capacitor circuits that would otherwise have no feedback connection at some instant of time. This capacitor would normally be small (around 0.5 pF or less).

This gain circuit can be either inverting or noninverting depending on the clock phases of the input stage. While the inverting circuit creates its output as a delay-free version of the input, the output for the noninverting case would be one-half clock cycle behind the input signal.

To see how this capacitive-reset gain circuit operates, consider the inverting circuit during the reset phase of ϕ_2, as shown in Fig. 14.35(*a*). We have assumed capacitor C_3 was charged to the output voltage during the previous ϕ_1 clock phase. Here we see that capacitors C_1 and C_2 are charged to the opamp's input-offset voltage, V_{off}, in the same manner as in the resettable gain circuit of Fig. 14.31. The next clock phase of ϕ_1 is shown in Fig. 14.35(*b*), where we see that the output voltage is independent of the offset voltage in the same manner as the resettable gain circuit. In addition, we see that during ϕ_1, capacitor C_3 is charged to the output voltage. Before proceeding, it is of interest to note what happens to the charges on C_1 and C_2 when going from phase ϕ_1 to ϕ_2 (i.e., circuit (*b*) to (*a*) of Fig. 14.35). Since the charges on C_1 and C_2 are equal (recall that the charge on C_2 was obtained during the charging of C_1), when one side of each of C_1 and C_2 are grounded, their charges cancel each other and no charge is dumped onto C_3. Therefore, in the circuit of Fig. 14.35(*a*), the voltage on C_3 remains equal to the previous output voltage as shown. It should be noted here that even if the two charges did not precisely cancel, it would only result in the output voltage moving slightly further away from the previous output voltage during ϕ_2 but would not affect the output voltage during ϕ_1.

Finally, this capacitive-reset gain circuit can also realize a differential-to-single-ended gain stage, as shown in Fig. 14.36. Note that the shown circuit indicates that the switches connected around the virtual grounds are disconnected slightly sooner than other switches to reduce nonlinearities due to charge injection. This circuit has a switched-capacitor circuit connected to the positive input of the opamp that not only accepts differential inputs but also, to a first-order approximation, cancels the clock feedthrough of the switches [Martin, 1982].

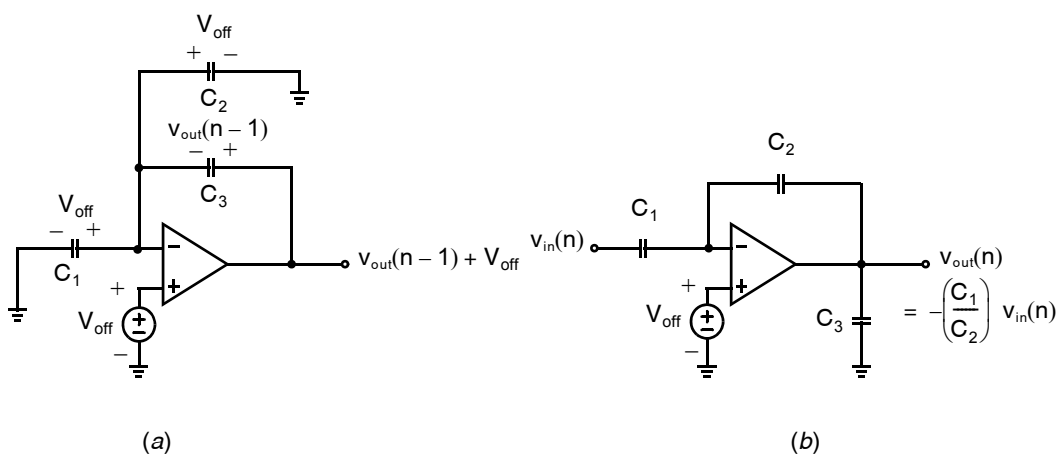


Fig. 14.35 Capacitive-reset gain circuit. (*a*) During reset (ϕ_2) when the output remains near previous output level. (*b*) During valid output (ϕ_1) when no offset error is present.

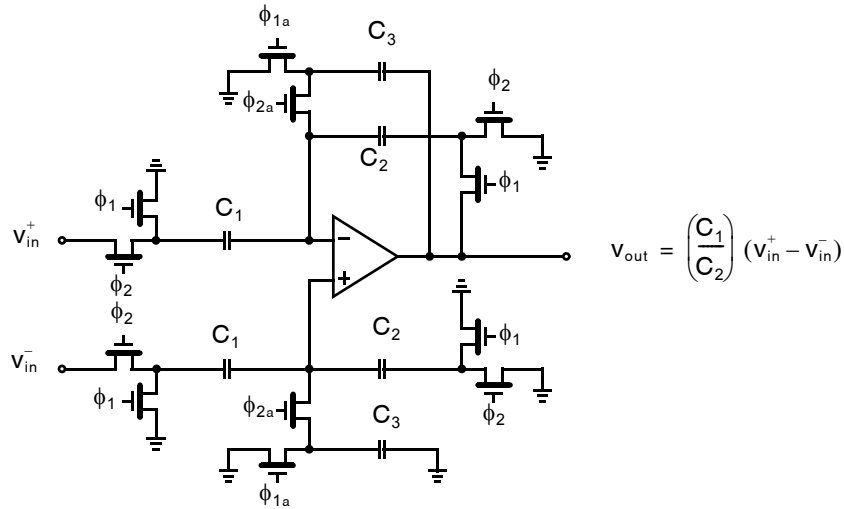


Fig. 14.36 A differential-to-single-ended gain circuit using capacitive reset.

14.8 CORRELATED DOUBLE-SAMPLING TECHNIQUES

The preceding SC gain amplifier is an example of using correlated double sampling (CDS) to minimize errors due to finite offset voltages, 1/f noise, and finite opamp gain. This technique is generally applicable to SC circuits of many different types [Temes, 1996]. It has been used to realize highly accurate gain amplifiers, sample and holds, and integrators. The basic methodology in all cases is similar: During a calibration phase, the finite input voltage of the opamp is sampled and stored across capacitors; during the operation phase (when the output is being sampled), this error voltage is subtracted from the signal voltage by appropriate switching of the capacitors. A detailed description of CDS techniques is beyond the scope of this text; rather, a couple of examples will be briefly described and the interested reader can consult the tutorial [Temes, 1996] for more information.

A wide-band amplifier that uses CDS, similar to the SC amplifier of Fig. 14.34, but with superior high-frequency operation, is shown in Fig. 14.37. During ϕ_2, C_1' and C_2' are used to have

$$v_{out} \cong -\frac{C_1'}{C_2'} v_{in} \tag{14.103}$$

but with errors due to finite input-offset voltage, 1/f noise, and gain. At the same time, the finite opamp input voltage caused by these errors is sampled and stored across C_1 and C_2. Next, during ϕ_1, this input error voltage is subtracted from the signal (applied to the opamp input) at that time. Assuming that the input voltage and the opamp input error voltages did not change appreciably from ϕ_2 to ϕ_1, errors due to them will be significantly reduced.

The technique can also be used to realize accurate integrators. For example, the SC integrator shown in Fig. 14.38 [Nagaraj, 1986; Ki, 1991] uses an additional capacitor, C_2', to sample the opamp input error voltages during ϕ_1. Next, during ϕ_2, C_2' is connected in series with the opamp's inverting input and greatly minimizes the effects of these errors (by a factor of the inverse of the opamp's gain over what would otherwise occur at frequencies substantially less than the sampling frequency). Similar architectures, with an open-loop capacitor in series with the opamp input during the phase when the output is sampled, can also be used for gain amplifiers.

When CDS sampling is used, the opamps should be designed to minimize thermal noise rather than 1/f noise. This might mean the use of n-channel input transistors rather than p-channel input transistors, for example.

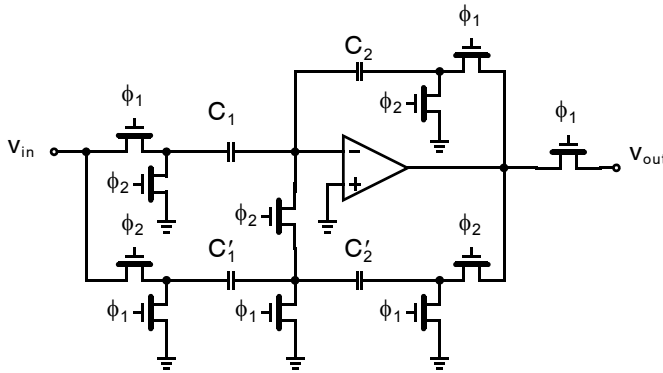


Fig. 14.37 An SC amplifier with CDS to minimize errors due to input- offset voltages, 1/f noise, and finite gain.

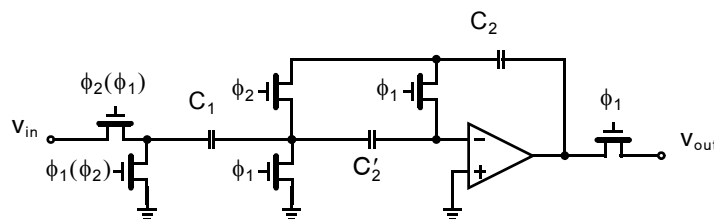


Fig. 14.38 An SC integrator with CDS to minimize errors due to input-offset voltages, 1/f noise, and finite gain.

When this technique is used in high-order SC filters, often only a couple of stages will have low-frequency gain from the opamp input terminals to the filter outputs and will require CDS circuitry. Other integrators can be more simply realized. This technique has proven to be very useful in applications such as oversampling A/D converters, where accurate integrators are required in the first stage [Hurst, 1989; Rebeschini, 1989] to reduce input-offset voltages and especially 1/f noise.

14.9 OTHER SWITCHED-CAPACITOR CIRCUITS

In this section, we present a variety of other switched-capacitor circuits useful for nonlinear applications. Specifically, we look at an amplitude modulator, full-wave rectifier, peak detector, voltage-controlled oscillator, and sinusoidal oscillator.

14.9.1 Amplitude Modulator

Amplitude modulators are used to shift a signal along the frequency axis. For example, to shift an information signal, $m(t)$, by a carrier frequency, ω_{ca}, we multiply $m(t)$ by a sinusoidal signal of frequency, ω_{ca}, or mathematically,

$$y(t) = m(t) \times \cos(\omega_{ca} t) \tag{14.104}$$

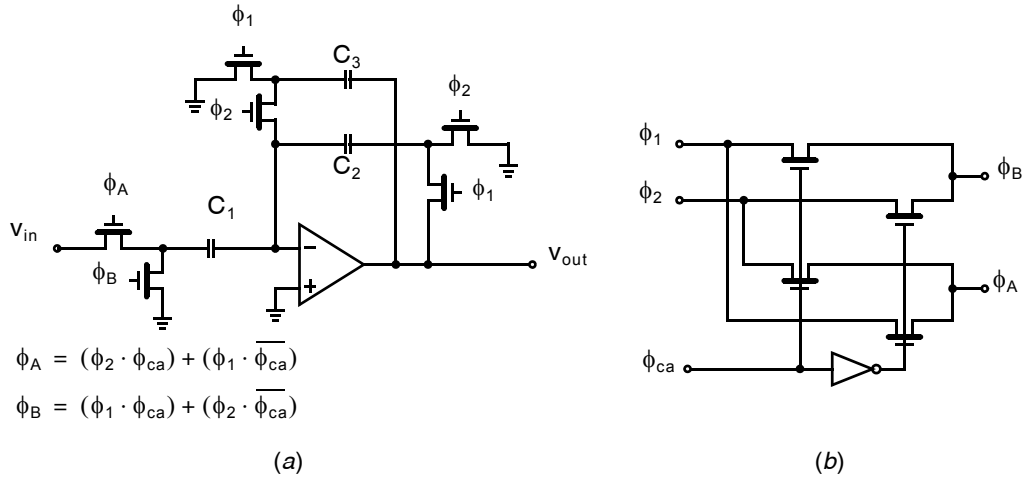


Fig. 14.39 (*a*) A switched-capacitor square-wave modulator where the input clock phases are controlled by the modulating square wave ϕ_M. (*b*) A possible circuit realization for ϕ_A and ϕ_B.

Since the Fourier transform of $\cos(\omega_{ca}t)$ is $\pi\delta(\omega - \omega_{ca}) + \pi\delta(\omega + \omega_{ca})$, and multiplication in the time domain is equivalent to convolution in the frequency domain, one can show that the spectrum of $y(t)$ is equal to

$$Y(\omega) = \frac{1}{2}M(\omega + \omega_{ca}) + \frac{1}{2}M(\omega - \omega_{ca}) \tag{14.105}$$

Thus, the output spectrum only has power around $\pm\omega_{ca}$.

However, in many cases it is difficult to realize a sinusoidal signal as well as a linear multiplier. As a result, many modulators make use of a square-wave carrier signal, in which case the output signal is simply

$$y(t) = m(t) \times s_q(t) = \pm m(t) \tag{14.106}$$

where $s_q(t)$ is a square wave operating at the carrier frequency whose amplitude has been normalized to 1. Thus, this approach simply requires a square-wave signal at the carrier frequency determining whether $m(t)$ or $-m(t)$ is passed to the output. The penalty paid here is that rather than $Y(\omega)$ having power located only around $\pm\omega_{ca}$, there will also be significant power around odd multiples of $\pm\omega_{ca}$ since a square-wave signal has power at odd harmonics of the fundamental. These extra images can typically be tolerated through the use of additional filtering stages.

A square-wave modulator can be realized using switched-capacitor techniques, as shown in Fig. 14.39. Here, the modulator is realized by making use of the capacitive-reset gain stage of Fig. 14.34. Note that the polarity of the output signal is determined by changing the clock phases of the input-transistor pair through the use of ϕ_A and ϕ_B, which are controlled by the square-wave carrier clock signal ϕ_{ca}. When ϕ_{ca} is high, $\phi_A = \phi_2$ and $\phi_B = \phi_1$, which results in a noninverting output of $(C_1/C_2)v_{in}$. When ϕ_{ca} goes low, the clock phases of ϕ_A, ϕ_B switch, and the circuit produces an inverting output of $-(C_1/C_2)v_{in}$.

14.9.2 Full-Wave Rectifier

A full-wave rectifier produces as its output the absolute value of the input. In other words, if the input is positive the output equals the input, but if the input is negative the output equals an inverted version of the input.

It is possible to realize a full-wave rectifier by making use of an amplitude modulator where, instead of a square-wave carrier signal being used, the output of a comparator controls the polarity of the output signal, as shown in Fig. 14.40. If the input signal is above ground, ϕ_{ca} goes high and the output is equal to a scaled version

of the input. If the input signal is below ground, ϕ_{ca} goes low and the output equals an inverted scaled version of input. For proper operation, the comparator output must change synchronously with the sampling instances. Circuits to achieve this operation are given in [Gregorian, 1986].

14.9.3 Peak Detectors

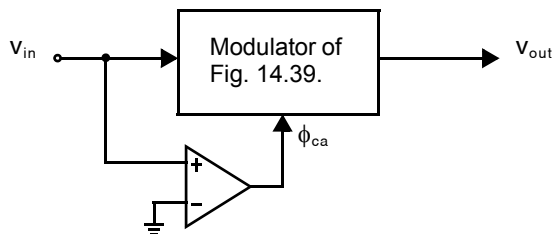


Fig. 14.40 A full-wave detector based on the square-wave modulator circuit of Fig. 14.39.

Peak detectors produce as their output the maximum value of an input signal over some time period. While peak detectors will normally have some decaying circuitry associated with them (or a circuit to reset the peak occasionally), here we shall just look at the peak detector circuit itself. Two methods for realizing peak detectors are shown in Fig. 14.41.

In Fig. 14.41(*a*), a latched comparator is used to determine when $v_{in} > v_{out}$. When such a condition occurs, the comparator output goes high and the switch Q_1 turns on, forcing $v_{out} = v_{in}$. The comparator is latched here to maintain its speed as well as ensuring that the comparator output is either high or low. Note that this circuit requires that the input signal directly drives C_H through Q_1, and hence buffering may be required. However, this circuit can be reasonably fast since latched comparators can be made to operate quickly and there is no opamp settling involved (only the RC time constant of Q_1 and C_H).

The circuit of Fig. 14.41(*b*) requires an opamp (with proper compensation), since it places Q_2 in a feedback connection. When $v_{in} > v_{out}$, the opamp output will start to go high, thus turning on Q_2 and hence turning on the feedback loop. While the feedback loop is on, the high loop gain will force $v_{out} = v_{in}$ with a high accuracy. Note that this circuit has a high input impedance as the input signal is applied to the noninverting terminal of an opamp. However, this opamp-based circuit will typically be slower than the comparator-based circuit of Fig. 14.41(*a*) since one needs to wait for the opamp output to slew and then settle. This slewing time can be minimized by adding circuitry to ensure that the gate of Q_2 never becomes lower than the output voltage when Q_2 is not conducting.

14.9.4 Voltage-Controlled Oscillator

A voltage-controlled oscillator (VCO) is an oscillator whose frequency can be adjusted through the use of a controlling voltage. As we shall see, the VCO described here generates an output square wave that places the oscillator

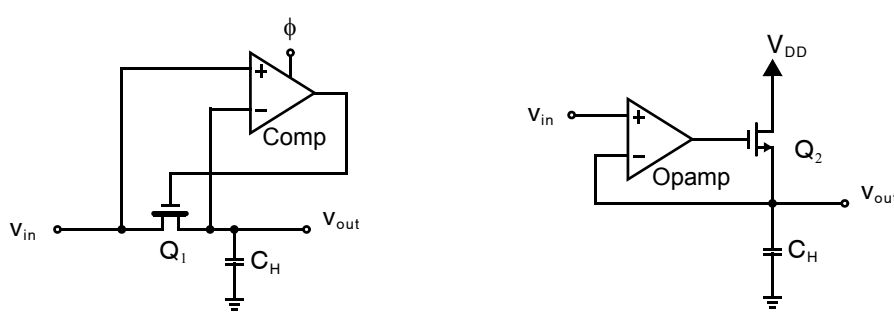


Fig. 14.41 Two-peak detectors: (*a*) latched-comparator approach; (*b*) continuous-time approach.

alternatively in one of two states and is referred to as a relaxation oscillator. As we shall see in Chapter 19, a VCO is an integral part of a phase-locked loop (PLL).

The basic principle and typical waveforms of the relaxation oscillator are shown in Fig. 14.42 [Martin, 1981] and operate as follows. Assume that v_{out} has just toggled low to $-V_{ss}$ and node v_x is now at a level quite a bit higher than zero. On ϕ_1, k_2C will be charged up to a voltage of $-V_{SS}$, resulting in a charge value given by

$$Q_{k_2C} = -k_2CV_{SS} \tag{14.107}$$

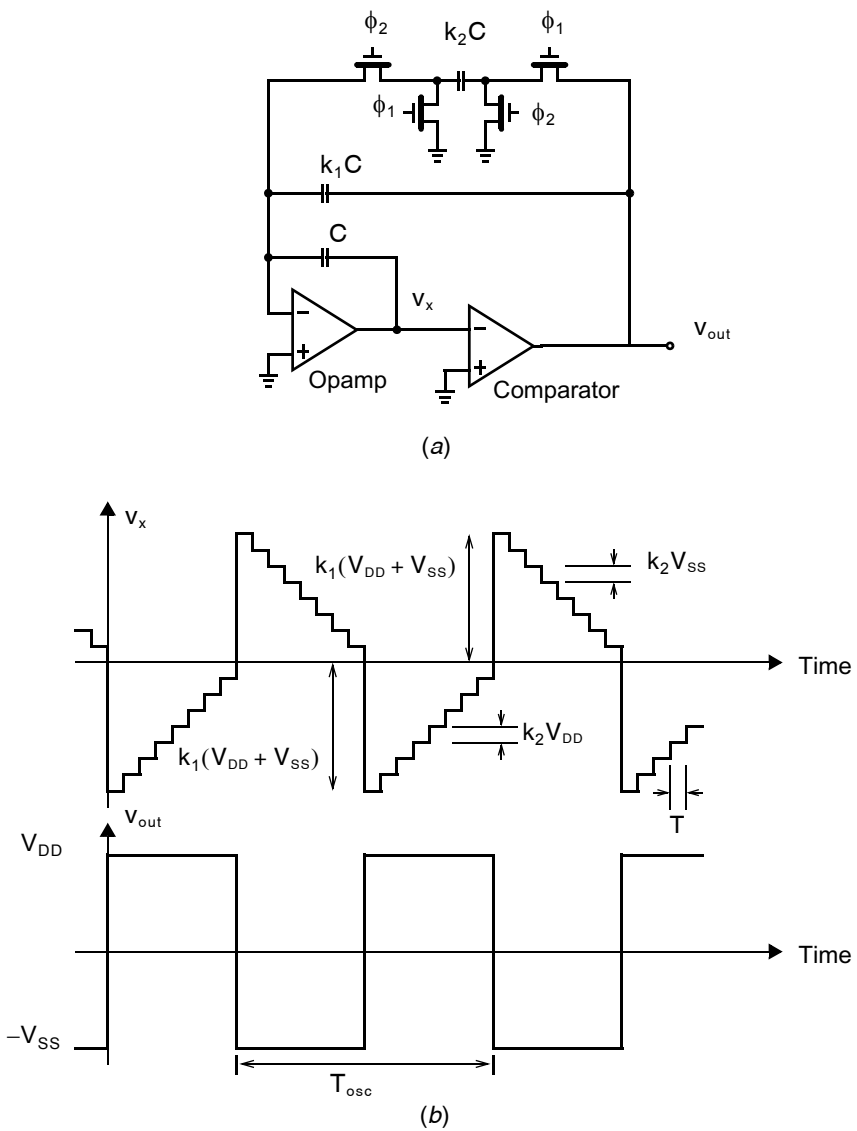


Fig. 14.42 A switched-capacitor relaxation oscillator.

Note that the voltage across the unswitched capacitor k_1C remains unchanged at $-V_{SS}$. Next, as ϕ_2 goes high, the charge on k_2C will be transferred to C (while the voltage across unswitched k_1C again remains unchanged) and the output will fall by an amount equal to

$$\Delta v_x = \frac{\Delta Q_C}{C} = \frac{Q_{k_2C}}{C} = -k_2 V_{SS} \tag{14.108}$$

This operation will continue until v_x reaches zero volts. In effect, k_2C and C have formed a noninverting discrete-time integrator, while capacitor k_1C does not affect the operation of the circuit. Once v_x goes below zero, the comparator output will toggle to V_{DD}, and at this point a large amount of charge will pass through k_1C while the voltage across it changes from $-V_{SS}$ to V_{DD}. This large charge value is equal to

$$\Delta Q_{k_1C} = k_1 C(V_{DD} - (-V_{SS})) \tag{14.109}$$

and since this charge is drawn across C, then $\Delta Q_C = \Delta Q_{k_1C}$ and node v_x drops by an amount equal to

$$\Delta v_x = -\frac{\Delta Q_C}{C} = -k_1(V_{DD} + V_{SS}) \tag{14.110}$$

as shown in Fig. 14.42(*b*). After this time, v_x will step up in increments of k_2V_{DD} (since v_{out} now equals V_{DD}) until $v_x = 0$, at which point a large positive voltage step will occur. Just after this positive step, we will be back at our starting point, and the entire operation will repeat indefinitely.

To derive a formula for the oscillation period T_{osc}, we make the assumption that $k_1 >> k_2$ such that many small steps occur over one-half T_{osc} period. With this assumption, we can now write that the total number of clock cycles, T_{osc}/T, equals the number of steps during negative sloping v_x plus the number of steps during positive sloping v_x. Mathematically, we have

$$\frac{T_{osc}}{T} = \frac{k_1(V_{DD} + V_{SS})}{k_2 V_{SS}} + \frac{k_1(V_{DD} + V_{SS})}{k_2 V_{DD}} \tag{14.111}$$

Assuming $V_{DD} = V_{SS}$, we have a 50 percent duty cycle with

$$T_{osc} = 4\left(\frac{k_1}{k_2}\right) T \tag{14.112}$$

or equivalently,

$$f_{osc} = \frac{1}{4}\left(\frac{k_2}{k_1}\right) f \tag{14.113}$$

This fixed-frequency oscillator can be made voltage controlled by adding an extra feed-in capacitor to the oscillator, $k_{in}C$, as shown in Fig. 14.43. Here, when V_{in} is positive, extra charge packets are added into the integrator with the same sign as those for k_2C, and the circuit reaches $v_x = 0$ sooner. If V_{in} is negative, then extra charge packets are added into the integrator with the opposite sign as those for k_2C, and the circuit reaches $v_x = 0$ later. Assuming $k_1 >> k_2, k_{in}$ and $V_{DD} = V_{SS}$, it is left as an exercise to the reader (Problem 14.25) to show that the frequency of oscillation for the voltage-controlled oscillator of Fig. 14.43 is given by

$$f_{osc} = \frac{1}{4}\left(\frac{k_2}{k_1} + \frac{k_{in}V_{in}}{k_1 V_{DD}}\right) f \tag{14.114}$$

14.9.5 Sinusoidal Oscillator

In many applications, it is useful to realize an oscillator which produces a nearly sinusoidal output. For example, when a button is pressed on a touch-tone telephone, the signal sent down the telephone line is simply the sum of

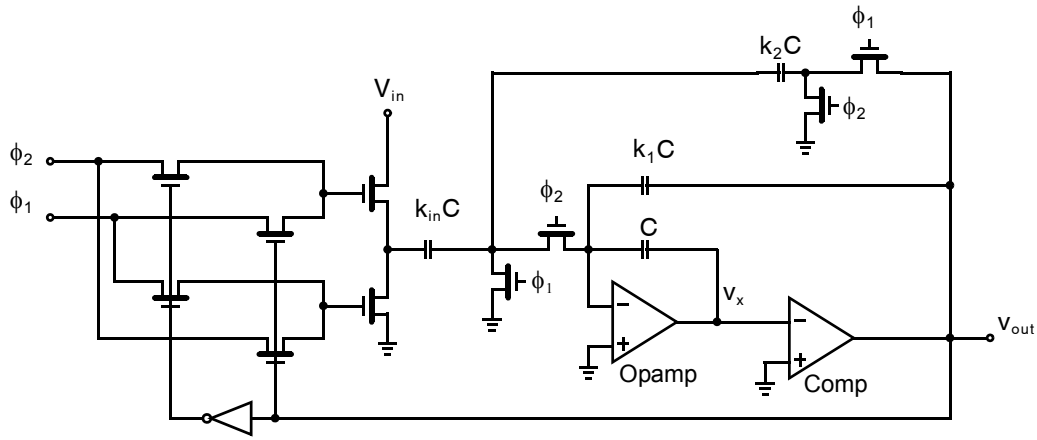


Fig. 14.43 A voltage-controlled oscillator (VCO) based on the relaxation oscillator of Fig. 14.42.

two sinusoidal signals at certain frequencies (the signal is known as a dual-tone multi-frequency (DTMF) signal). Thus, switched-capacitor sinusoidal oscillators could be used to generate the necessary touch-tone signals.

One method of generating a sinusoidal oscillator is shown in Fig. 14.44. Here, a high-Q bandpass filter with center frequency f_0 is used together with a hard limiter. Assuming some start-up circuit is used to force oscillations to exist, we see that a sinusoidal wave at v_{out} results in a square wave at v_1 which has a large component at f_0. The signal v_1 is filtered by the bandpass filter and thus maintains the sinusoidal signal, v_{out}. The oscillation frequency is set by the center frequency of the bandpass filter, while the sinusoidal amplitude is set by the center frequency gain of the filter and the peak values of the square-wave signal.

A switched-capacitor implementation of this sinusoidal oscillator is shown in Fig. 14.45 [Fleischer, 1985], where the high-Q bandpass filter has been realized using the circuit of Fig. 14.28. The comparator generates signals X and $\overline{X}$ with the result that the switched-input associated with V_{ref} is either positive or negative. The start-up circuit is used to sense when an ac signal is present at v_{out}. When no ac signal is detected, v_{st} is low and the extra capacitors C_5 and C_6 form a positive feedback loop which causes oscillations to build up. Once an ac signal is detected, v_{st} goes high and the positive feedback loop is broken, thereby allowing oscillations to continue, as discussed previously.

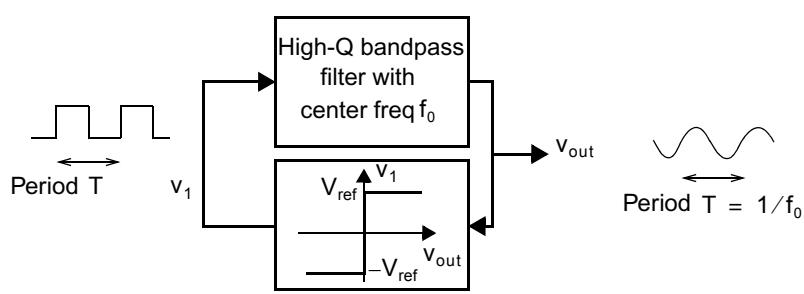


Fig. 14.44 A sinusoidal oscillator that makes use of a bandpass filter.

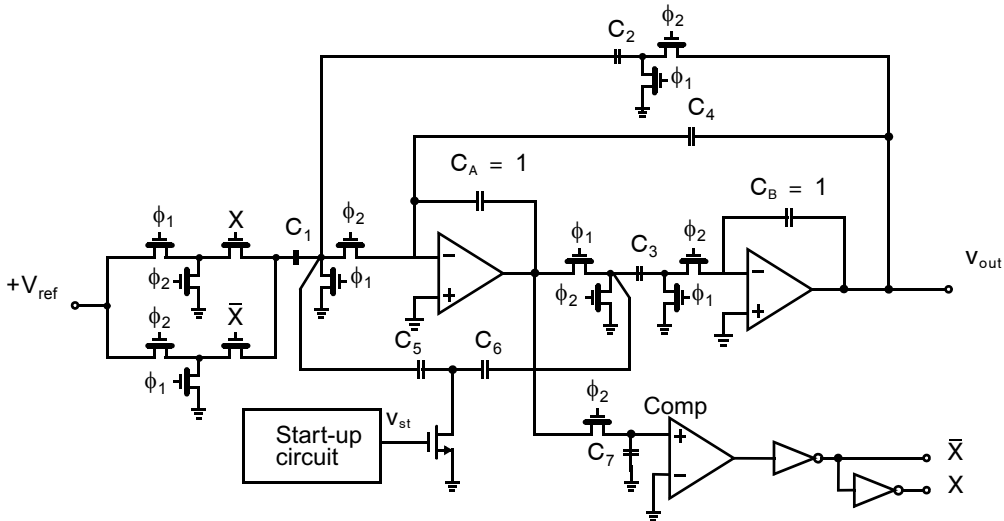


Fig. 14.45 A switched-capacitor implementation of a sinusoidal oscillator.

For examples of other nonfiltering uses of switched-capacitor circuits, the interested reader is referred to [Gregorian, 1986].

14.10 KEY POINTS

- Opamp nonidealities such as dc gain, unity-gain frequency and phase-margin, slew-rate, and dc offset impact switched-capacitor circuit performance. [p. 557]
- When signal voltages change very slowly compared with the clock frequency, a switched-capacitor gives rise to an average current inversely proportional to the voltage across it, behaviorally equivalent to a resistor. [p. 561]
- When signal voltages change appreciably from one clock cycle to the next, a discrete-time analysis is needed to accurately predict the input–output behavior of switched-capacitor circuits. [p. 561]
- Each clock cycle, a switched-capacitor integrator samples a quantity of charge proportional to an input voltage and the sampling capacitance, then delivers it to a second integrating capacitor. [p. 562]
- The discrete-time transfer function of a switched-capacitor integrator has a pole at dc, z = 1. [p. 563]
- In order to make a switched-capacitor integrator insensitive to the circuit's parasitic capacitances to ground, it is necessary to add two extra switches, for a total of four, on each switched capacitor. [p. 567]
- Both delayed and delay-free integrators can be realized with the same switched-capacitor circuit by changing the clock phases applied to two of the switches. In a fully-differential implementation, both can be configured to provide either positive or negative dc gain. [p. 568]
- Complicated switched-capacitor circuits may be analyzed graphically using a signal-flow-graph representing the opamp stage with an integrating block and replacing each input branch with the corresponding non switched, delayed, or delay-free gain factor. [p. 570]
- Switched capacitors accumulate the thermal noise of the switches. Sampled kT/C noise from both clock phases superimpose resulting in a total mean squared input-referred noise of 2kT/C. [p. 570]
- Beginning with conventional continuous-time active RC filters, and replacing each resistor with a switched-capacitor results in a discrete-time filter whose frequency response closely approximates that of the original continuous-time filter at frequencies far below the sampling frequency. Once a filter structure is obtained, its precise frequency response is determined through the use of discrete-time analysis. [p. 572]

- Fully-differential filter implementations offer several advantages over their single-ended counterparts: they allow for the realization of sign inversion (i.e. a gain of –1) by simply interchanging wires, reject common-mode noise, and rejecting even-order distortion terms which generally dominate single-ended circuits. [p. 576]
- Although fully-differential circuit implementations require some overhead for additional wiring and common-mode feedback circuitry, if properly designed the overhead should represent a small fraction of the overall circuit. [p. 577]
- Several different discrete-time biquads can be realized by varying the type (i.e., delayed, delay-free) and exact connection of switched-capacitor branches within a two-integrator loop. The choice of which biquad to use depends upon the specific pole and zero locations sought. [p. 584]
- To minimize distortion due to charge injection, switches connected to dc nodes, such as ground, are turned off in advance of other switches. Although there is still some charge injection, it is signal independent. [p. 586]
- Charge injection is most problematic at high switching speeds where low switch on resistances demand wide transistors and small capacitors. [p. 587]
- Technology scaling permits smaller transistors to offer the same on resistance, thereby reducing charge injection or, for the same charge injection, enabling higher switching frequencies. [p. 587]
- Switched-capacitor circuits can use a reset phase to sample and cancel dc offset and very low-frequency (1/f) noise without affecting the circuit's overall dc gain. [p. 589]

14.11 REFERENCES

D. Allstot, R. Brodersen, and P. R. Gray. "MOS Switched-Capacitor Ladder Filters," *IEEE J. of Solid-State Circuits,* Vol. SC-13, no. 6, pp. 806–814, December 1978.

P. O. Brackett and A. S. Sedra. *Filter Theory and Design: Active and Passive.* Matrix Publishers, Champaign, Illinois, p. 575, 1978.

L. T. Bruton. "Low-Sensitivity Digital Ladder Filters," *IEEE Trans. Circuits and Systems,* Vol. CAS-22, no. 3, pp. 168–176, March 1975.

J. T. Caves, M. A. Copeland, C. F. Rahim, and S. D. Rosenbaum. "Sampled-Data Filters Using Switched Capacitors as Resistor Equivalents," *IEEE J. of Solid-State Circuits,* Vol. SC-12, pp. 592–600, December 1977.

P. E. Fleischer, A. Ganesan, and K. R. Laker, "A Switched-Capacitor Oscillator with Precision Amplitude Control and Guaranteed Start-up," *IEEE J. of Solid-State Circuits,* Vol. SC-20, pp. 641–647, April 1985.

D. L. Fried. "Analog Sample-Data Filters," *IEEE J. of Solid-State Circuits,* Vol. SC-7, pp. 302–304, August 1972.

T. Foxall, R. Whitbread, L. Sellars, A. Aitken, and J. Morris. "A Switched-Capacitor Bandsplit Filter Using Double Polysilicon Oxide-Isolated CMOS," *1980 ISSCC Dig. of Technical Papers,* pp. 90–91, February 1980.

R. Gregorian. "High-Resolution Switched-Capacitor D/A Converter," *Microelectronics J.,* Vol. 12, no. 2, pp. 10–13, March/April 1981.

R. Gregorian, K. Martin, and G. C. Temes. "Switched-Capacitor Circuit Design," *IEEE Proceedings,* Vol. 71, no. 8, pp. 941–966, August 1983.

R. Gregorian and G. C. Temes. *Analog MOS Integrated Circuits for Signal Processing,* John Wiley & Sons, New York, 1986.

B. J. Hosticka, R. W. Broderson, and P. R. Gray. "MOS Sampled Data Recursive Filters Using Switched-Capacitor Integrators," *IEEE J. of Solid-State Circuits,* Vol. SC-12, no. 6, pp. 600–608, December 1977.

P. J. Hurst and R. A. Levinson. "Delta-Sigma A/D with Reduced Sensitivity to Op-Amp Noise and Gain," *Proc. of IEEE Int. Symp. on Circuits and Systems,* pp. 254–257, 1989.

G. M. Jacobs, D. J. Allstot, R. W. Brodersen, and P. R. Gray. "MOS Switched-Capacitor Ladder Filters," *IEEE Trans. Circuits and Systems,* Vol. CAS-25, pp. 1014–1021, December 1978.

W. H. Ki and G. C. Temes. "Offset- and Clock-Feedthrough-Compensated SC Filters," *Proc. of IEEE Int. Symp. on Circuits and Systems,* pp. 1561–1564, 1991.

K. R. Laker and W. M. C. Sansen. *Design of Analog Integrated Circuits and Systems*. McGraw-Hill, New York, 1994.

L. E. Larson and G. C. Temes. "SC Building Blocks with Reduced Sensitivity to Finite Amplifier Gain, Bandwidth, and Offset Voltage," *Proc. of IEEE Int. Symp. on Circuits and Systems,* pp. 334–338, 1987.

K. Martin. "Improved Circuits for the Realization of Switched-Capacitor Filters," *IEEE Trans. Circuits and Systems*, Vol. CAS-27, no. 4, pp. 237–244, April 1980; Also published as BNR Tech. Rep. TR1E 81-78-06, March 1978.

K. Martin. "Improved Circuits for the Realization of Switched Capacitor Filters," *Proc. 1979 Int. Symp. on Circuits and Systems (IEEE & IECE),* Tokyo, pp. 756–760, July 1979.

K. Martin. "A Voltage-Controlled Switched-Capacitor Relaxation Oscillator," *IEEE J. of Solid State Circuits,* Vol. SC-16, no. 4, pp. 412–414, August 1981.

K. Martin. "New Clock-Feedthrough Cancellation Technique for Analogue MOS Switched-Capacitor Circuits," *Electron. Lett.,* Vol. 18, no. 1, pp. 39–40, January 1982.

K. Martin. L. Ozcolak, Y. S. Lee, and G. C. Temes. "A Differential Switched-Capacitor Amplifier," *IEEE J. of Solid-State Circuits,* Vol. SC-22, no. 1, pp. 104–106, February 1987.

K. Martin and S. Rosenbaum. "Improvements in Sampled Analog Filter," filed March 1978, in Canada, the United States and Japan. Issued in Canada October 21, 1980, patent no. 1088161. Issued in England June 10, 1982, no. CYB 2019151B.

K. Martin and A. S. Sedra. "Strays-Insensitive Switched-Capacitor Filters Based on Bilinear z-Transform," *Electron. Lett.,* Vol. 15, no. 13, pp. 365–366, June 1979.

K. Martin and A. S. Sedra. "Exact Design of Switched-Capacitor Bandpass Filters Using Coupled-Biquad Structures," *IEEE Trans. Circuits and Systems*, Vol. CAS-27, no. 6, pp. 469–478, June 1980.

K. Martin and A. S. Sedra. "Switched-Capacitor Building Blocks for Adaptive Systems," *IEEE Trans. Circuits and Systems,* Vol. CAS-28, no. 6, pp. 576–584, June 1981.

K. Nagaraj et al. "SC Circuits with Reduced Sensitivity to Amplifier Gain," *Proc. of IEEE Int. Symp. on Circuits and Systems*, pp. 618–621, 1986.

H. Matsumoto and K. Watanabe. "Spike-Free Switched-Capacitor Circuits," *Electron. Lett.,* Vol. 23, no. 8, pp. 428–429, April 1987.

W. Poschenrieder, "Frequencz Filterung durch Netzwerke mit Periodis Gesteurten Schaltern," *Analys und Synthese von N etzwerken,* pp. 220–237, Tanungsheft, Stuttgart, 1966.

M. Rebeschini et al. "A High-Resolution CMOS Sigma-Delta A/D Converter with 320 kHz Output Rate," *Proc. of IEEE Int. Symp. on Circuits and Systems,* pp. 246–249, 1989.

R. Schaumann, M. S. Ghausi, and K. R. Laker. *Design of Analog Filters*. Prentice Hall, Englewood Cliffs, New Jersey, 1990.

R. Schreier, J. Silva, J. Steensgaard, and G. C. Temes, "Design-Oriented Estimation of Thermal Noise in Switched-Capacitor Circuits," *IEEE Trans. Circuits and Systems - I: Regular Papers*, vol. 42, no. 11, pp. 2358-2368, November 2005.

G. C. Temes and C. Enz. "Circuit techniques for reducing the effects of op-amp imperfections: Autozeroing, correlated double sampling, and chopper stabilization," *IEEE Proceedings,* vol.84, no. 11, pp. 1584-1614, November 1996.

B. J. White, G. M. Jacobs, and G. F. Landsburg. "A Monolithic Dual-Tone Multifrequency Receiver," *IEEE J. of Solid-State Circuits,* Vol. SC-14, no. 6, pp. 991–997, December 1978.

I. A. Young, P. R. Gray, and D. A. Hodges. "Analog NMOS Sampled-Data Recursive Filters," *IEEE Int. Symp. on Solid-State Circuits,* pp. 156–157, February 1977.

14.12 PROBLEMS

14.12.1 Section 14.1: Basic Building Blocks

14.1 Assuming logic signals have infinitely fast fall and rise times but are delayed by 1 ns through a logic gate, draw the output waveforms of the two phase clock generators shown in Fig. 14.3(*b*). Assume the delay blocks add delays of 10 ns each and a clock frequency of 10 MHz is applied to the input.

14.2 You must estimate the size of NMOS transistors required to implement the switches in a switched-capacitor circuit. Use the transistor parameters for the 0.35 μm technology in Table 1.5 and the minimum gate length. Assume the gate voltage is at $V_{DD} = 3.3$ V and both drain and source are at $V_{DD}/2$. Also estimate the total gate capacitance ($C_{gs} + C_{gd}$) of the transistor when it is on. Ensure the on resistance is 5 kΩ

14.3 Repeat Problem 14.2 using the 0.18 μm and 45 nm technologies in Table 1.5 with $V_{DD} = 1.8$ V and 1.0 V respectively. In each case, assume the minimum gate length is used, the gate voltage is V_{DD}, and both source and drain are at $V_{DD}/2$. How do the results compare between technologies?

14.12.2 Section 14.2: Basic Operation and Analysis

14.4 Ignoring the effect of parasitic capacitances, find the discrete-time transfer function of the switched-capacitor circuit shown in Fig. P14.4.

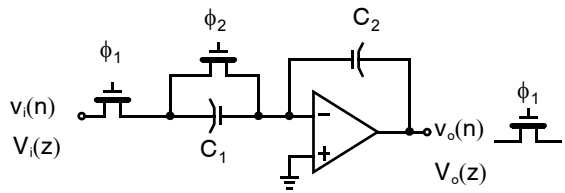


Fig. P14.4

14.5 Repeat Problem 14.4 but include the effect of parasitic capacitances.

14.6 Show that when an opamp has a finite gain of A, the transfer function for the discrete-time integrator of Fig. 14.5 is given by

$$\frac{V_o(z)}{V_i(z)} = -\left(\frac{C_1}{C_2}\right)\left(\frac{1}{z\left(1+\dfrac{C_1}{C_2 A}\right)-1}\right)$$

Also show that this transfer function has a dc gain of $-A$ and a pole that is located slightly to the left of 1 by an amount approximately equal to $(C_1/C_2)\,(1/A)$.

14.7 Find the transfer function and a single-ended circuit implementation of the signal-flow-graph in Fig. P14.7. Find expressions for the pole- and zero-locations.

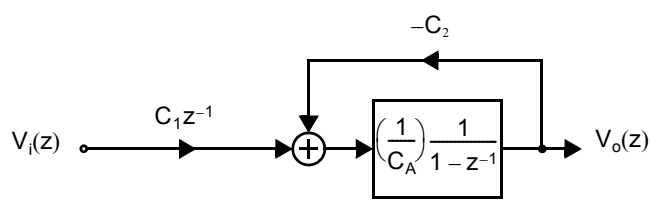


Fig. P14.7

14.2.3 Section 14.3: Noise in Switched-Capacitor Circuits

14.8 A switched-capacitor circuit is to be designed with an SNR of 60 dB for input sinusoidal signals of 1 V peak-to-peak. Find the required value of the switched capacitance.

14.9 Size a switched capacitor to provide an input-referred rms noise of less than 0.4 mV. Then, find the width of the switch transistors to provide a sampling time constant, $2R_{on}C$, of 1 ns. Use the transistor parameters for the 0.18 μm technology in Table 1.5 and assume that the NMOS switches have a V_{eff} of at least 300 mV.

14.10 Show that the input referred noise of a delay-free switched-capacitor branch is the same as that of the delayed switched-capacitor analyzed in Section 14.3.

14.12.4 Section 14.4: First-Order Filters

14.11 Repeat Example 14.3 but assume the discrete-time zero is at $z = 0$ and the pole remains at $z = 0.53327$. Note that with this change, C_1 is no longer required at the sacrifice of a nonzero gain being obtained at 50 kHz. What is the new gain at 50 kHz?

14.12 Find the capacitances needed for the first-order filter shown in Fig. 14.18(*a*) to realize a discrete-time transfer function having a dc gain of one, a zero at 0, and a 3-dB frequency of 1 kHz when a clock frequency of 50 kHz is used. Assume that $C_A = 50$ pF. What is the gain (in dB) at 25 kHz?

14.13 Find the transfer function of the circuit shown in Fig. 14.19 when $C_1 = 0$ pF, $C_2 = 2$ pF, $C_3 = 2$ pF, and $C_A = 20$ pF. What is the magnitude and phase of the gain at dc, $f_s/4$ and $f_s/2$?

14.14 Find the discrete-time transfer function of the switched capacitor circuit shown in Fig. P14.14.

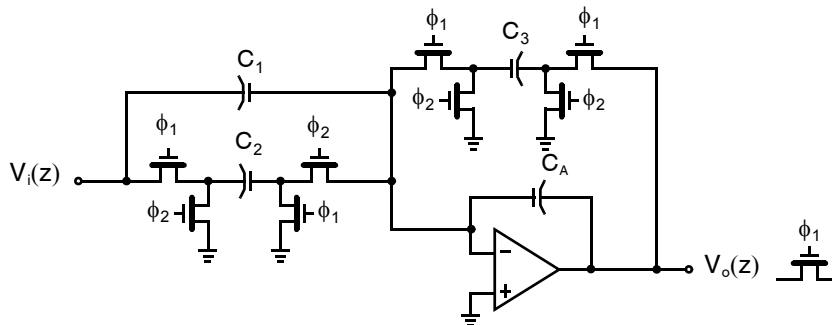


Fig. P14.14

14.15 Draw the schematic of a fully-differential implementation of the switched capacitor circuit shown in Fig. P14.14 where it is desired to realize a negative value for the capacitance C_2.

14.12.5 Section 14.5: Biquad Filters

14.16 Show that when $K_6 = 0$ in the biquad circuit of Fig. 14.24, if the poles are complex, they lie precisely on the unit circle (i.e., the circuit is a resonator).

14.17 Select the capacitor ratios, K_i, in the biquad circuit of Fig. 14.24 so that the dc gain is unity, both poles lie at $z_p = 0.8$, and zeros lie at $z_z = \pm j$, resulting in a notch in the frequency response at $f_s/4$. Assume a fully-differential implementation so that negative capacitor ratios are possible.

14.18 Using the bilinear transform, find the transfer function in the z-domain of a discrete-time bandpass filter with a $Q = 20$ (in the continuous-time domain) and a peak gain of one near $f_s/100$. Find the largest to smallest capacitor ratio if this transfer function is realized using the high-Q biquad circuit, and compare it to that which would be obtained if the low-Q circuit were used instead. Let $C_1 = C_2 = 1$ in both cases.

14.12.6 Section 14.6: Charge Injection

14.19 Which of the clock signals in Fig. P14.14 should be advanced in order to reduce the effects of charge injection?

14.20 The circuit of Fig. 14.29 with $C_1 = 0$ and $C_2 = C_A = 5C_3 = 5$ pF is realized with the 0.18 μm technology in Table 1.5. Switches are NMOS devices with a gate voltage $V_{DD} = 1.8$ V and the schematic's "ground" voltage located $V_{DD}/2 = 0.9$ V above the opamp's lower supply. Find the size of switches $Q_{3,4}$ required to ensure a dc offset of less than 5 mV due to charge injection. If all switches are the same size, and allowing 5 time constants for complete settling, what is the resulting settling time of switched capacitor C_2?

Problems 14.20 and 14.21 can be simulated using an ideal opamp subcircuit.

14.21 Repeat Problem 14.20, but this time size the switches to ensure a settling time of 40 ns. Find the resulting dc offset due to charge injection.

14.12.7 Section 14.7: Switched-Capacitor Gain Circuits

14.22 Sketch a typical output waveform for the capacitive-reset gain circuit of Fig. 14.34 if the opamp has an input offset of 100 mV and a gain of 10 is desired.

14.23 Consider the capacitive-reset gain circuit during each of the two clock phases depicted in Fig. 14.35. What is the feedback factor, β, in each of those two configurations? Which clock phase will, therefore, result in the worst-case phase margin for the feedback circuit?

14.12.8 Section 14.8: Correlated Double-Sampling Techniques and Section 14.9: Other Switched-Capacitor Circuits

14.24 Draw the circuit of Fig. 14.37 in each of its two clock phases with on switches replaced by short circuits and off switches replaced by open circuits. Repeat for the circuit of Fig. 14.38.

14.25 Verify that Eq. (14.114) gives the frequency of oscillation for the VCO of Fig. 14.43.

CHAPTER

15 Data Converter Fundamentals

In this chapter, fundamental aspects of analog-to-digital (A/D) and digital-to-analog (D/A) converters are presented without regard for their internal architecture or circuit design. In effect, converters are treated in this chapter as black boxes such that only their input–output relationships are discussed. Internal architectures and circuits techniques for realizing data converters are discussed in following chapters.

Before proceeding, it is useful to make the distinction between two main types of data converters.

Key Point: *In Nyquist-rate converters, there is a one-to-one correspondence between input and output values. Oversampling converters increase SNR by operating much faster than the input signal's Nyquist rate and filtering out quantization noise outside of the signal bandwidth.*

Nyquist-Rate Converters We loosely define *Nyquist-rate data converters* as those converters that generate a series of output values in which each value has a one-to-one correspondence with a single input value. For example, a Nyquist-rate D/A converter would generate a series of analog output levels, where each level is a result of a single B-bit input word. However, it should be noted that Nyquist-rate converters are seldom used at the Nyquist rate due to the difficulty in realizing practical anti-aliasing and reconstruction filters. In most cases, Nyquist rate converters operate at 1.5 to 10 times the Nyquist rate (i.e., 3 to 20 times the input signal's bandwidth).

Oversampling Converters *Oversampling converters* are those converters that operate much faster than the input signal's Nyquist rate (typically 10 to 512 times faster) and increase the output's signal-to-noise ratio (SNR) by filtering out quantization noise that is not in the signal's bandwidth. In A/D converters, this filtering is performed digitally, whereas in D/A converters, analog filtering is used. Most often, oversampling converters use *noise shaping* to place much of the quantization noise outside the input signal's bandwidth (see Chapter 18).

This distinction is made here since much of the descriptions that follow refer more closely to Nyquist-rate converters than to oversampling converters.

15.1 IDEAL D/A CONVERTER

Consider the block diagram of an N-bit D/A converter shown in Fig. 15.1. Here, B_{in} is defined to be an N-bit digital signal (or word) such that

$$B_{in} = b_1 2^{-1} + b_2 2^{-2} + \cdots + b_N 2^{-N} \tag{15.1}$$

where b_i equals 1 or 0 (i.e., b_i is a binary digit). We also define b_1 as the most significant bit (MSB) and b_N as the least significant bit (LSB). Furthermore, we have assumed here that B_{in} represents a positive value, resulting in a unipolar D/A converter. A *unipolar D/A converter* produces an output signal of only one polarity. In contrast, *signed converters* produce output signals of either positive or negative polarity, depending on a sign bit (usually b_1). Extending the following concepts to the signed case is straightforward but requires knowledge of the type of digital representation used (i.e., sign magnitude, offset binary, or 2's complement). Signed codes are discussed in Section 15.4.

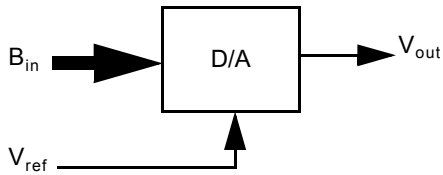


Fig. 15.1 A block diagram representing a D/A converter.

The analog output signal, V_{out}, is related to the digital signal, B_{in}, through an analog reference signal, V_{ref}. For simplicity, we assume that both V_{out} and V_{ref} are voltage signals, although, in general, they may be other physical quantities, such as current or charge. The relationship between these three signals for a unipolar D/A converter is given by

$$V_{out} = V_{ref}(b_1 2^{-1} + b_2 2^{-2} + \cdots + b_N 2^{-N}) = V_{ref} B_{in} \qquad (15.2)$$

It is useful to define V_{LSB} to be the voltage change when one LSB changes, or, mathematically,

$$V_{LSB} \equiv \frac{V_{ref}}{2^N} \qquad (15.3)$$

Also useful (particularly in measuring errors) is the definition of a new "unit," namely, LSB units, which are in fact unitless.

$$1 \text{ LSB} = \frac{1}{2^N} \qquad (15.4)$$

The transfer curve for an ideal 2-bit D/A converter is shown in Fig. 15.2. Note here that, although only a finite number of analog values occur at the output, for an ideal D/A converter, the output signals are well-defined values. Also note that the maximum value of V_{out} is not V_{ref} but rather $V_{ref}(1 - 2^{-N})$, or equivalently, $V_{ref} - V_{LSB}$. Finally, as we can see from (15.2), a multiplying D/A converter is realized by simply allowing the reference signal, V_{ref}, to be a varying input signal along with the digital input, B_{in}. Such an arrangement results in V_{out} being proportional to the multiplication of the input signals, B_{in} and V_{ref}.

Key Point: *Data converters translate between digital binary codes and analog voltages or currents. In an ideal converter, the binary codes correspond to analog signal quantities that are precisely equally spaced.*

EXAMPLE 15.1

An 8-bit D/A converter has $V_{ref} = 5$ V. What is the output voltage when $B_{in} = 10110100$? Also, find V_{LSB}.

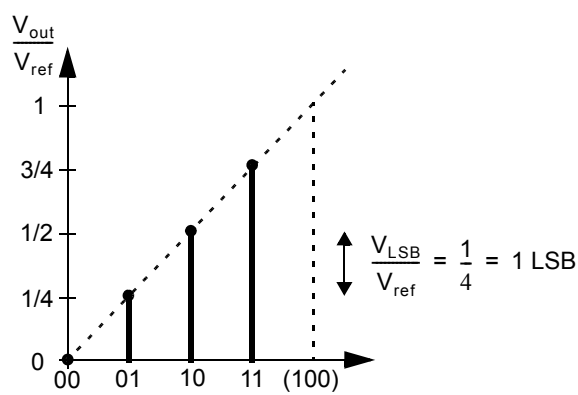


Fig. 15.2 Input–output transfer curve for an ideal 2-bit D/A converter.

Solution

We can find the decimal equivalent of B_{in} using (15.1)

$$B_{in} = 2^{-1} + 2^{-3} + 2^{-4} + 2^{-6} = 0.703125 \tag{15.5}$$

Then, using (15.2), we find

$$V_{out} = V_{ref}B_{in} = 3.516 \text{ V} \tag{15.6}$$

and

$$V_{LSB} = 5/256 = 19.5 \text{ mV} \tag{15.7}$$

15.2 IDEAL A/D CONVERTER

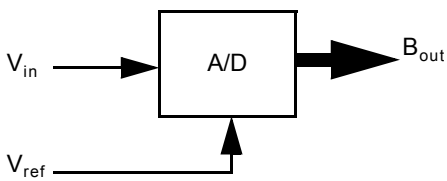


Fig. 15.3 A block diagram representing an A/D converter.

The block diagram representation for an A/D converter is shown in Fig. 15.3, where B_{out} is the digital output word while V_{in} and V_{ref} are the analog input and reference signals, respectively. Also, we define V_{LSB} to be the signal change corresponding to a single LSB change as in the D/A case.

For an A/D converter, the following equation relates these signals,

$$V_{ref}(b_1 2^{-1} + b_2 2^{-2} + \cdots + b_N 2^{-N}) = V_{in} \pm V_x$$

where

$$-\frac{1}{2}V_{LSB} \le V_x < \frac{1}{2}V_{LSB} \tag{15.8}$$

Note that there is now a range of valid input values that produce the same digital output word. This signal ambiguity produces what is known as *quantization error.* Also, note that no quantization error occurs in the case of a D/A converter since the output signals are well defined.[1]

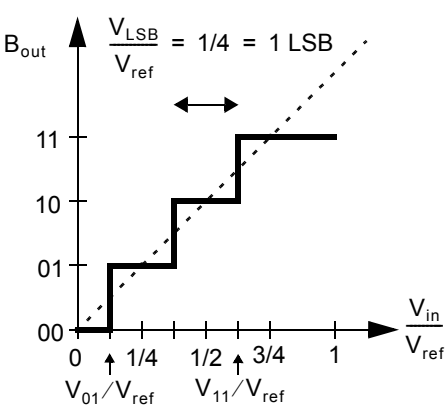


Fig. 15.4 Input–output transfer curve for a 2-bit A/D converter.

As in the D/A case, a transfer curve for an A/D converter can be sketched as shown in Fig. 15.4 for a 2-bit converter. Note that the transitions along the V_{in} axis are offset by $1/2\ V_{LSB}$, so that the midpoints of the staircase curve fall precisely on the equivalent D/A transfer curve. Here, we define the transition voltages at V_{ij}, where the subscript ij indicates the upper B_{out} value of the transition. For example, V_{01} is shown in Fig. 15.4 as the voltage (normalized with respect to V_{ref}) for the transition from 00 to 01.

Finally, it should be noted that the relation shown in (15.8) holds only if the input signal remains within 1 LSB of the two last transition voltages. Specifically, for the 2-bit transfer curve

1. Of course, quantization errors will occur if a 6-bit D/A converter is used to convert a 10-bit digital signal, but in this case, the quantization error occurs in the conversion from a 10-bit digital signal to a 6-bit digital signal.

shown in Fig. 15.4, V_{in} should remain less than $7/8\ V_{ref}$ and greater than $-1/8\ V_{ref}$. Otherwise, the quantizer is said to be *overloaded* since the magnitude of the quantization error would be larger than $V_{LSB}/2$.

15.3 QUANTIZATION NOISE

As mentioned in Section 15.2, quantization errors occur even in ideal A/D converters. In this section, we model these errors as being equivalent to an additive noise source and then find the power of this noise source. Consider the setup shown in Fig. 15.5, where both N-bit converters are ideal.

Since we have

$$V_Q = V_I - V_{in} \tag{15.9}$$

we can rearrange this equation as

$$V_I = V_{in} + V_Q \tag{15.10}$$

Although this rearrangement is trivial, it has important implications because (15.10) shows that the quantized signal, V_I, can be modelled as the input signal, V_{in}, plus some additive quantization noise signal, V_Q. Note that (15.10) is exact because no approximations have been made here. The quantization noise modelling becomes approximate once some assumptions are made about the statistical properties of V_Q.

15.3.1 Deterministic Approach

To gain an understanding of some of the properties of the quantization noise signal, V_Q, we will investigate its behavior when the input is a particular function. Specifically, let us assume the input signal, V_{in}, in Fig. 15.5 is a ramp. Such an input signal results in the output from the D/A, V_I, appearing as a staircase, as shown in Fig. 15.6, assuming no overloading occurs. Taking the difference between these two signals gives us the noise signal, V_Q, which is a

Key Point: *As long as it is not overloaded, the ideal quantizer introduces an additive quantization noise less than 1/2 of an LSB in magnitude.*

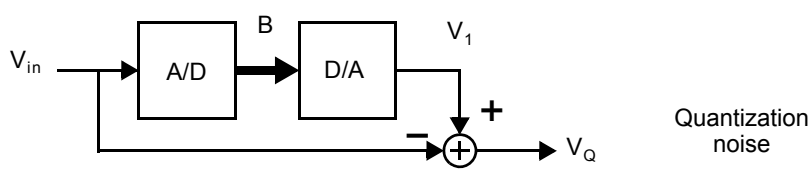


Fig. 15.5 A circuit to investigate quantization noise behavior.

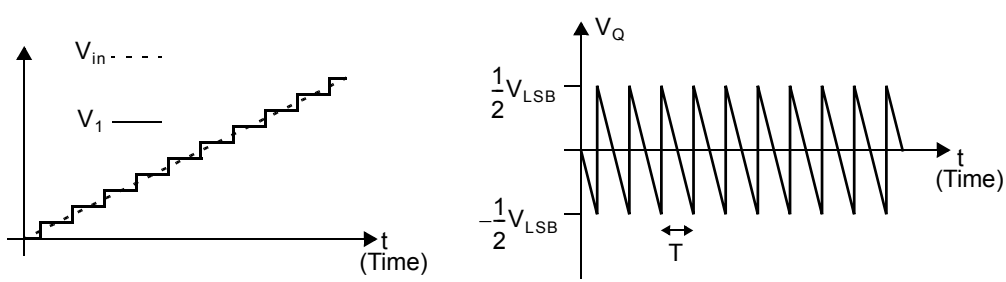


Fig. 15.6 Applying a ramp signal to the circuit in Fig. 15.5.

result of quantization error. Note that the quantization signal, V_Q, is limited to $\pm V_{LSB}/2$ and will be so limited for all input signals (not just ramps). Clearly, the average of V_Q is zero. However, the rms value of the noise signal, $V_{Q(rms)}$, is given by

$$V_{Q(rms)} = \left[\frac{1}{T}\int_{-T/2}^{T/2} V_Q^2\, dt\right]^{1/2} = \left[\frac{1}{T}\int_{-T/2}^{T/2} V_{LSB}^2\left(\frac{t}{T}\right)^2 dt\right]^{1/2}$$
$$= \left[\frac{V_{LSB}^2}{T^3}\left(\frac{t^3}{3}\Big|_{-T/2}^{T/2}\right)\right]^{1/2} \tag{15.11}$$

$$V_{Q(rms)} = \frac{V_{LSB}}{\sqrt{12}} \tag{15.12}$$

Thus, we see that the rms power of the quantization noise source is proportional to the size of V_{LSB}, which is determined by the number of bits, N, in the converter.

15.3.2 Stochastic Approach

The preceding deterministic approach was presented as a simple example to see some properties of the quantization noise signal. However, to deal with the more general input case, a stochastic approach is typically used. In a stochastic approach, we assume that the input signal is varying rapidly such that the quantization error signal, V_Q, is a random variable uniformly distributed between $\pm V_{LSB}/2$. The probability density function for such an error signal, $f_Q(x)$, will be a constant value, as shown in Fig. 15.7.

The average value of the quantization error, $V_{Q(avg)}$, is found to be zero as follows:

$$V_{Q(avg)} = \int_{-\infty}^{\infty} x f_Q(x)\, dx = \frac{1}{V_{LSB}}\left(\int_{-V_{LSB}/2}^{V_{LSB}/2} x\, dx\right) = 0 \tag{15.13}$$

In a similar fashion, the rms value of the quantization error is given by

$$V_{Q(rms)} = \left[\int_{-\infty}^{\infty} x^2 f_e(x)\, dx\right]^{1/2} = \left[\frac{1}{V_{LSB}}\left(\int_{-V_{LSB}/2}^{V_{LSB}/2} x^2\, dx\right)\right]^{1/2} = \frac{V_{LSB}}{\sqrt{12}} \tag{15.14}$$

which is the same result as (15.12), which is found using the deterministic ramp input signal. The fact that these two results are identical should come as no surprise since randomly chosen samples from the sawtooth waveform in the deterministic case would also have a uniformly distributed probability density function. In general, the rms quantization noise power equals $V_{LSB}/\sqrt{12}$ when the quantization noise signal is uniformly distributed over the interval $\pm V_{LSB}/2$.

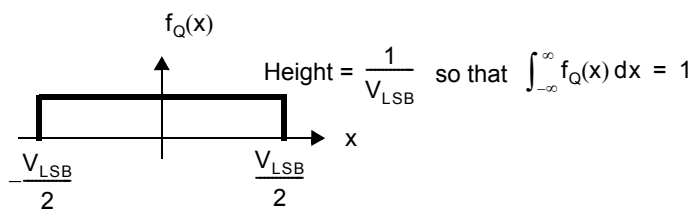


Fig. 15.7 Assumed probability density function for the quantization error, V_Q.

Recalling that the size of V_{LSB} is halved for each additional bit and assuming that V_{ref} remains constant, we see from (15.14) that the noise power decreases by 6 dB for each additional bit in the A/D converter. Thus, given an input signal waveform, a formula can be derived giving the best possible signal-to-noise ratio (SNR) for a given number of bits in an ideal A/D converter.

For example, assuming V_{in} is a sawtooth of height V_{ref} (or equivalently, a random signal uniformly distributed between 0 and V_{ref}) and considering only the ac power of the signal, the signal-to-quantization noise ratio (SQNR) is given by

$$\begin{aligned} \text{SQNR} &= 20\log\left(\frac{V_{in(rms)}}{V_{Q(rms)}}\right) = 20\log\left(\frac{V_{ref}/\sqrt{12}}{V_{LSB}/\sqrt{12}}\right) \\ &= 20\log(2^N) = 6.02N \text{ dB} \end{aligned} \tag{15.15}$$

This represents the best possible SNR achievable for a data converter with N-bits of resolution since only the quantization noise of an ideal converter is considered. For example, a 10-bit A/D converter has a best possible SNR of about 60 dB.

Alternatively, a more common SNR formula is to assume V_{in} is a sinusoidal waveform between 0 and V_{ref}. Thus, the ac power of the sinusoidal wave is $V_{ref}/(2\sqrt{2})$, which results in

$$\begin{aligned} \text{SQNR} &= 20\log\left(\frac{V_{in(rms)}}{V_{Q(rms)}}\right) \\ &= 20\log\left(\frac{V_{ref}/(2\sqrt{2})}{V_{LSB}/(\sqrt{12})}\right) \\ &= 20\log\left(\sqrt{\frac{3}{2}}2^N\right) \\ \text{SQNR} &= 6.02N + 1.76 \text{ dB} \end{aligned} \tag{15.16}$$

In other words, a sinusoidal signal has 1.76 dB more ac power than a random signal uniformly distributed between the same peak levels.

Note that (15.16) gives the best possible SNR for an N-bit A/D converter. However, the idealized SNR decreases from this best possible value for reduced input signal levels. For example, Fig. 15.8 shows a plot of the idealized SNR for a 10-bit A/D converter versus the sinusoidal input signal amplitude. However, it should be noted that these SNR values could be improved through the use of oversampling techniques if the input signal's bandwidth is lower than the Nyquist rate. Oversampling will be discussed in detail in Chapter 18, where the design of oversampling converters is presented.

Key Point: *Quantization noise may be modeled as a random quantity uniformly distributed between –LSB/2 and +LSB/2. For a sinusoidal input, this results in a maximum SNR of 6.02N + 1.76 dB for an ideal N-bit converter.*

EXAMPLE 15.2

A 200-mV_{pp} sinusoidal signal is applied to an ideal 12-bit A/D converter for which $V_{ref} = 5$ V. Find the SNR of the digitized output signal.

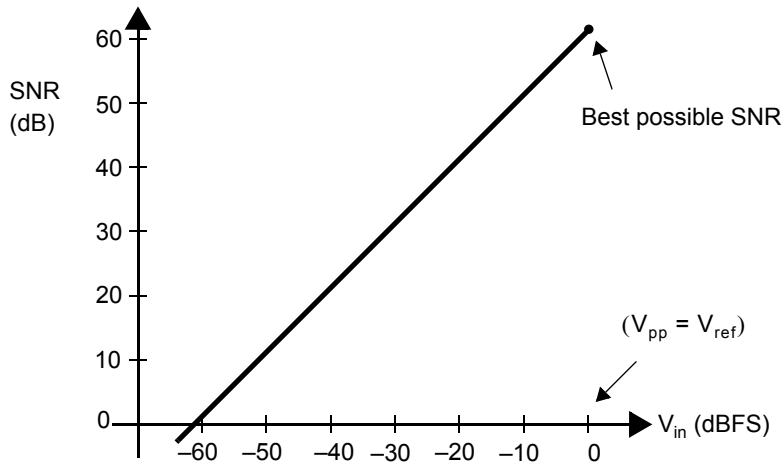


Fig. 15.8 Idealized SNR versus sinusoidal input signal amplitude for a 10-bit A/D converter. The 0-dBFS input signal amplitude corresponds to a peak-to-peak voltage equaling "Full-Scale," in this case V_{ref}.

Solution

First, we use (15.16) to find the maximum SNR if a full-scale sinusoidal waveform of ±2.5 V were applied to the input.

$$\text{SNR}_{max} = 6.02 \times 12 + 1.76 = 74 \text{ dB} \tag{15.17}$$

However, since the input is only a ±100-mV sinusoidal waveform that is 28 dB below full scale, the SNR of the digitized output is

$$\text{SNR} = 74 - 28 = 46 \text{ dB} \tag{15.18}$$

15.4 SIGNED CODES

In many applications, it is necessary to create a converter that operates with both positive and negative analog signals, resulting in the need for both positive and negative digital representations. Typically, the analog signal is bounded by $\pm 0.5 V_{ref}$, such that its full-scale range is the same magnitude as in the unipolar case. Some common signed digital representations are sign magnitude, 1's complement, offset binary, and 2's complement, as shown in Table 15.1 for the 4-bit case. Note from Table 15.1 that all positive number representations are the same except for the offset-binary case, where the MSB is complemented.

Sign Magnitude

For negative numbers in the sign-magnitude case, all the bits are the same as for the positive number representation, except that the MSB is complemented. For example, the sign-magnitude representation for 5 is 0101, whereas for –5 it is 1101. However, note that this approach results in two representations for the number 0, and thus only $2^N - 1$ numbers are represented.

Table 15.1 Some 4-bit signed digital representations.

Number	Normalized number	Sign magnitude	1's complement	Offset binary	2's complement
+7	+7/8	0111	0111	1111	0111
+6	+6/8	0110	0110	1110	0110
+5	+5/8	0101	0101	1101	0101
+4	+4/8	0100	0100	1100	0100
+3	+3/8	0011	0011	1011	0011
+2	+2/8	0010	0010	1010	0010
+1	+1/8	0001	0001	1001	0001
+0	+0	0000	0000	1000	0000
(–0)	(–0)	(1000)	(1111)		
–1	–1/8	1001	1110	0111	1111
–2	–2/8	1010	1101	0110	1110
–3	–3/8	1011	1100	0101	1101
–4	–4/8	1100	1011	0100	1100
–5	–5/8	1101	1010	0011	1011
–6	–6/8	1110	1001	0010	1010
–7	–7/8	1111	1000	0001	1001
–8	–8/8			0000	1000

1's Complement

In 1's-complement representation, negative numbers are represented as the complement of all the bits for the equivalent positive number. For example, here, 5 is once again 0101, whereas –5 is now 1010. It should be noted that the 1's-complement case also has two representations for 0, and thus only $2^N - 1$ numbers are represented.

Offset Binary

The offset-binary representation is obtained by assigning 0000 to the most negative number and then counting up, as in the unipolar case. In other words, this system can be thought of as simply a unipolar representation counting from 0 to 2^N, but where the decimal numbers represented are offset by 2^{N-1}, or equivalently, the decimal counts are from -2^{N-1} to 2^{N-1}. For example, the offset-binary code for the number 5 in the 4-bit case is the same as the unipolar code for the number 13, which is 1101, since $13 = 5 + 2^{4-1}$. The offset-binary code for –5 is the same as the unipolar code for 3, which is 0011. Note that the offset-binary code does not suffer from redundancy, and all sixteen numbers are uniquely represented. Also, this code has the advantage that it is closely related to the unipolar case through a simple offset.

Finally, the unipolar relationship for a D/A converter given in (15.2) is easily modified to the signed case, with offset-binary representation as

$$V_{out} = V_{ref}(b_1 2^{-1} + b_2 2^{-2} + \cdots + b_N 2^{-N}) - 0.5V_{ref} \tag{15.19}$$

Note here that the output signal is now bounded by $\pm 0.5V_{ref}$.

2's Complement

Finally, the 2's-complement representation is obtained from the offset-binary number by simply complementing the MSB. For the 4-bit example, 5 becomes 0101 in 2's complement (the same as in sign magnitude and in 1's complement), whereas –5 is now 1011. It should be mentioned that the 2's-complement code for negative numbers can also be obtained by adding 1 LSB to the equivalent 1's-complement code.

The main advantage of 2's-complement coding is that addition of both positive and negative numbers is performed using straightforward addition, and no extra hardware is required. Also, the subtraction of two numbers, $A - B$, can be easily performed by complementing all the bits for B (i.e., forming the 1's-complement equivalent) and then adding this result to A at the same time as adding a single LSB in order to create the 2's-complement equivalent of $-B$. Adding a single LSB is easily accomplished in hardware by setting the *carry-in bit* high in the overall adder. Thus, subtraction requires only a small amount of extra hardware. Finally, if many numbers are being added using 2's-complement codes, no overflow hardware is required as long as the final result is within the digital code range (even if intermediate results go well out of range). For these reasons, 2's-complement codes are the most popular representation for signed numbers when arithmetic operations are to be performed.

EXAMPLE 15.3

Show that, when adding $2/8$, $7/8$, and $-3/8$ together as 2's-complement numbers, the final result is correct although an intermediate result is incorrect.

Solution

The addition of $2/8$ and $7/8$ is given by

$$0010 + 0111 = 1001 \tag{15.20}$$

which corresponds to $-7/8$ (note that this temporary result is two steps past $7/8$ when thinking of a 2's-complement code as being circular). Although this intermediate result is incorrect since overflow occurred, we simply carry on and add $-3/8$ to find the final answer.

$$1001 + 1101 = 0110 \tag{15.21}$$

Thus, the final result of $6/8$ is correct, although a temporary overflow occurred, as we just saw.

15.5 PERFORMANCE LIMITATIONS

In this section, some commonly used terms describing the performance of data converters are defined. Before proceeding, definitions are required for determining the transfer responses of both D/A and A/D converters. The transfer response of a D/A converter is defined to be the analog levels that occur for each of the digital input words. Similarly, the transfer response of an A/D converter can be defined as the mid-points of the quantization intervals for each of the digital output words. However, since transitions are easier to measure than midpoint values, A/D converter errors are often measured in terms of the analog transition point values, V_{ij}, and we use that approach here.

15.5.1 Resolution

The *resolution* of a converter is defined to be the number of distinct analog levels corresponding to the different digital words. Thus, an N-bit resolution implies that the converter can resolve 2^N distinct analog levels. Resolution is not necessarily an indication of the accuracy of the converter, but instead it usually refers to the number of digital input or output bits.

15.5.2 Offset and Gain Error

In a D/A converter, the *offset error*, E_{off}, is defined to be the output that occurs for the input code that should produce zero output, or mathematically,

$$E_{off(D/A)} = \left.\frac{V_{out}}{V_{LSB}}\right|_{0\ldots0} \tag{15.22}$$

where the offset error is in units of LSBs. Similarly, for an A/D converter, the offset error is defined as the deviation of $V_{0\ldots01}$ from $1/2$ LSB, or mathematically,

$$E_{off(A/D)} = \frac{V_{0\ldots01}}{V_{LSB}} - \frac{1}{2}\ \text{LSB} \tag{15.23}$$

The gain error is defined to be the difference at the full-scale value between the ideal and actual curves when the offset error has been reduced to zero. For a D/A converter, the gain error, $E_{gain(D/A)}$, in units of LSBs, is given by

$$E_{gain(D/A)} = \left(\left.\frac{V_{out}}{V_{LSB}}\right|_{1\ldots1} - \left.\frac{V_{out}}{V_{LSB}}\right|_{0\ldots0}\right) - (2^N - 1) \tag{15.24}$$

For an A/D converter, the equivalent gain error, $E_{gain(A/D)}$ (in units of LSBs), is given by

$$E_{gain(A/D)} = \left(\frac{V_{1\ldots1}}{V_{LSB}} - \frac{V_{0\ldots01}}{V_{LSB}}\right) - (2^N - 2) \tag{15.25}$$

Graphical illustrations of gain and offset errors are shown in Fig. 15.9.

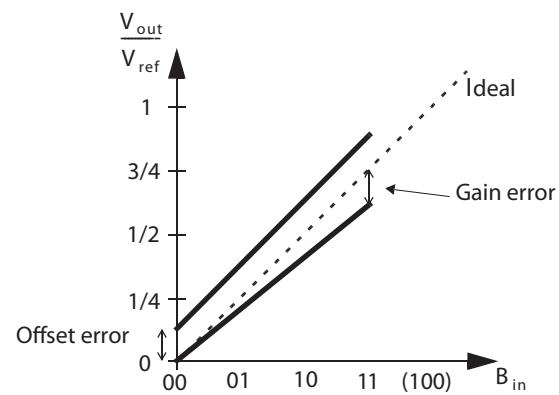


Fig. 15.9 Illustrating offset and gain err ors for a 2-bit D/A converter.

15.5.3 Accuracy and Linearity

The *absolute accuracy* of a converter is defined to be the difference between the expected and actual transfer responses. The absolute accuracy includes the offset, gain, and linearity errors.

The term *relative accuracy* is sometimes used and is defined to be the accuracy after the offset and gain errors have been removed. It is also referred to as the *maximum integral nonlinearity error* (described shortly) and we will refer to it as such.

Accuracy can be expressed as a percentage error of full-scale value, as the effective number of bits, or as a fraction of an LSB. For example, a 12-bit accuracy implies that the converter's error is less than the full-scale value divided by 2^{12}.

Note that a converter may have 12-bit resolution with only 10-bit accuracy, or 10-bit resolution with 12-bit accuracy. An accuracy greater than the resolution means that the converter's transfer response is very precisely controlled (better than the number of bits of resolution).

Key Point: *Integral nonlinearity (INL) is a popular measure of accuracy that specifies the deviation of a converter's input–output relationship away from the ideal, or least-squares fit, linear relationship.*

Integral Nonlinearity (INL) Error After both the offset and gain errors have been removed, the *integral nonlinearity (INL) error* is defined to be the deviation from a straight line. However, what straight line should be used? A conservative measure of nonlinearity is to use the endpoints of the converter's transfer response to define the straight line. An alternative definition is to find the best-fit straight line such that the maximum difference (or perhaps the mean squared error) is minimized. These two definitions are illustrated in Fig. 15.10. One should be aware that, in this book, we define INL values for each digital word (and thus these values can be plotted for a single converter), whereas others sometimes define the term "INL" as the maximum magnitude of the INL values (or equivalently, as the relative accuracy). Typically INL is measured by slowly sweeping the converter input over its full-scale range; hence it is a measure of the converter's accuracy only at low input signal frequencies.

Key Point: *Both DNL and INL are measured at dc or very low input frequencies. They are therefore referred to as measures of static nonlinearity.*

Differential Nonlinearity (DNL) Error In an ideal converter, each analog step size is equal to 1 LSB. In other words, in a D/A converter, each output level is 1 LSB from adjacent levels, whereas in an A/D, the transition values are precisely 1 LSB apart. *Differential nonlinearity* (DNL) is defined as the variation in analog step sizes away from 1 LSB (typically, once gain and offset errors have been removed). Thus, an ideal converter has its maximum differential nonlinearity of 0 for all digital values, whereas a converter with a maximum differential nonlinearity of 0.5 LSB has its step sizes varying from 0.5 LSB to 1.5 LSB. Once again, as in the INL case, we define DNL values for each digital word, whereas others sometimes refer to DNL as the maximum magnitude of the DNL values. Like INL, DNL is a measure of dc or low-frequency accuracy. Hence, both are referred to as measures of *static nonlinearities*.

Monotonicity A *monotonic* D/A converter is one in which the output always increases as the input increases. In other words, the slope of the D/A converter's transfer response is of only one sign. If the maximum DNL error is less than 1 LSB, then a D/A converter is guaranteed to be monotonic. However, many monotonic converters may have a maximum DNL greater than 1 LSB. Similarly, a converter is guaranteed to be monotonic if the maximum INL is less than 0.5 LSB.

Missing Codes Although monotonicity is appropriate for D/A converters, the equivalent term for A/D converters is *missing codes*. An A/D converter is guaranteed not to have any missing codes if the maximum DNL error is less than 1 LSB or if the maximum INL error is less than 0.5 LSB.

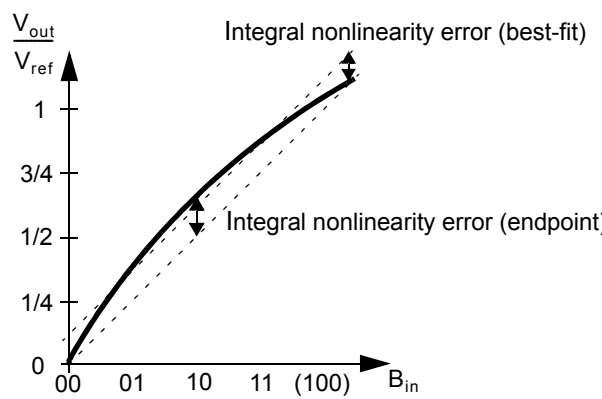


Fig. 15.10 Integral nonlinearity error in a 2-bit D/A converter.

A/D Conversion Time and Sampling Rate In an A/D converter, the *conversion time* is the time taken for the converter to complete a single measurement including acquisition time of the input signal. On the other hand, the maximum sampling rate is the speed at which samples can be continuously converted and is typically the inverse of the conversion time. However, one should be aware that some converters have a large latency between the input and the output due to pipelining or multiplexing, yet they still maintain a high sampling rate. For example, a pipelined 12-bit A/D converter may have a conversion time of 2 ns (i.e., a sampling rate of 500 MHz) yet a latency from input to output of 24 ns.

D/A Settling Time and Sampling Rate In a D/A converter, the *settling time* is defined as the time it takes for the converter to settle to within some specified amount of the final value (usually 0.5 LSB). The *sampling rate* is the rate at which samples can be continuously converted and is typically the inverse of the settling time.

Sampling-Time Uncertainty Both A/D and D/A converters have limited accuracy when their sampling instances are ill defined. To quantify this *sampling time uncertainty*, also known as *aperture jitter*, for sinusoidal waveforms, consider a full-scale signal, V_{in}, applied to an N-bit, signed, A/D converter with frequency f_{in}. Mathematically,

$$V_{in} = \frac{V_{ref}}{2} \sin(2\pi f_{in} t) \tag{15.26}$$

Since the rate of change (or slope) of V_{in} at the peak of a sinusoidal waveform is small, sampling time uncertainty is less of a problem near the peak values. However, the maximum rate of change for this waveform occurs at the zero crossing and can be found by differentiating V_{in} with respect to time and setting $t = 0$. At the zero crossing, we find that

$$\left.\frac{\Delta V}{\Delta t}\right|_{max} = \pi f_{in} V_{ref} \tag{15.27}$$

If Δt represents some sampling-time uncertainty, and if we want to keep ΔV less than 1 V_{LSB}, we see that

$$\Delta t < \frac{V_{LSB}}{\pi f_{in} V_{ref}} = \frac{1}{2^N \pi f_{in}} \tag{15.28}$$

For example, an 8-bit converter sampling a full-scale 250-MHz sinusoidal signal must keep its sampling-time uncertainty under 5 ps to maintain 8-bit accuracy. Also, the same 5-ps time accuracy is required for a 16-bit converter operating on a full-scale, 1-MHz signal.

Key Point: *Uncertainty in the sampling times of a converter, called aperture jitter, can be a significant performance limitation at high input signal frequencies.*

Dynamic Range In general terms, the *dynamic range* of a converter is the ratio of the maximum to minimum signal amplitudes that it can meaningfully process. A popular metric quantifying dynamic range is the maximum achievable signal-to-noise and distortion ratio (SNDR) specified as the ratio of the rms value of an input sinusoidal signal to the rms output noise plus the distortion measured when that same sinusoid is present at the output. In order to obtain the rms output noise plus distortion, the sinusoidal signal must first be eliminated from the measured output. In a D/A converter, the output sinusoid can be eliminated by using a spectrum analyzer and ignoring the power at that particular frequency. For an A/D converter, a similar approach can be taken by using an FFT and eliminating the fundamental of the output, or a least-mean-squared fit can be used to find the amplitude and phase of a sinusoid at the input signal's frequency and then subtracting the best-fit sinusoid from the output signal. Dynamic range can also be expressed as an effective number of bits (N_{eff}) by computing the resolution of an ideal

converter whose SQNR presented in (15.16) is exactly the same as the SNDR of the converter under test. Specifically, assuming SNDR is measured with a single-tone sinusoidal input,

$$N_{eff} = \frac{SNDR - 1.76\ dB}{6.02} \quad \text{bits} \tag{15.29}$$

Substitution of a measured SNDR into (15.29) will generally result in a fractional number of bits (for example, a converter may have 8.7 effective bits) which is perfectly fine since N_{eff} is merely a performance metric relating a converter's performance to that of a fictitious ideal quantizer. Note that if measured properly, SNDR includes quantization noise and it therefore can not exceed the SQNR in (15.16), so that $N_{eff} < N$ for any N-bit converter.

Key Point: *The most common measure of accuracy at high input signal frequencies is dynamic range, often quantified by the maximum achievable signal-to-noise and distortion ratio (SNDR). This may be expressed in dB, or as an effective number of bits. A converter's effective resolution bandwidth specifies the band over which SNDR is within 3 dBof its highest value.*

It should be noted that this approach often results in a dynamic range measurement that is a function of the frequency of the sinusoidal input. Hence, it is a more realistic way of measuring the performance of a converter than extrapolating the nonlinearity performance found using dc inputs. For example, if an 8-bit, 200-MHz, A/D converter has a band-limited preamplifier or a slew-rate limited sample and hold, dc inputs may show full 8-bit performance even at the maximum sample rate of 200 Msample/s. However, a high-frequency, sinusoidal input, at, say, 40 MHz, will require the input stage to track a rapidly varying signal and may result in only 6-bit performance. To capture this limitation, the *efffective resolution bandwidth* is specified as the bandwidth over which a converter's peak SNDR is within 3 dB of its best value, or equivalently the bandwidth over which N_{eff} is within 0.5 bits of its peak value.

Finally, it should be mentioned that the distortion level (or nonlinearity performance) of some converters remains at a fixed level and is not a function of the input signal level. Thus, as the input signal level is decreased, the signal-to-distortion ratio decreases. This behavior occurs because the distortion level is often determined by component matching and thus is fixed once the converter is realized. However, in some converters, the distortion level decreases as the input signal level is decreased, which is often a desirable property to have. For example, most 1-bit oversampling converters have this desirable property since they do not rely on component matching and their distortion level is often a result of weak nonlinear effects at the input stage.

EXAMPLE 15.4

Consider a 3-bit D/A converter in which $V_{ref} = 4$ V, with the following measured voltage values:

$$\{0.011 : 0.507 : 1.002 : 1.501 : 1.996 : 2.495 : 2.996 : 3.491\}$$

1. Find the offset and gain errors in units of LSBs.
2. Find the INL (endpoint) and DNL errors (in units of LSBs).
3. Find the effective number of bits of absolute accuracy.
4. Find the effective number of bits of relative accuracy.

Solution

We first note that 1 LSB corresponds to $V_{ref}/2^3 = 0.5$ V .

1. Since the offset voltage is 11 mV, and since 0.5 V corresponds to 1 LSB, we see that the offset error is given by

$$E_{off(D/A)} = \frac{0.011}{0.5} = 0.022 \text{ LSB} \tag{15.30}$$

For the gain error, from (15.25) we have

$$E_{\text{gain(D/A)}} = \left(\frac{3.491 - 0.011}{0.5}\right) - (2^3 - 1) = -0.04 \text{ LSB} \tag{15.31}$$

2. For INL and DNL errors, we first need to remove both offset and gain errors in the measured D/A values. The offset error is removed by subtracting 0.022 LSB off each value, whereas the gain error is eliminated by subtracting off scaled values of the gain error. For example, the new value for 1.002 (scaled to 1 LSB) is given by

$$\frac{1.002}{0.5} - 0.022 + \left(\frac{2}{7}\right)(0.04) = 1.993 \tag{15.32}$$

Thus, the offset-free, gain-free, scaled values are given by

$$\{0.0 : 0.998 : 1.993 : 2.997 : 3.993 : 4.997 : 6.004 : 7.0\}$$

Since these results are in units of LSBs, we calculate the INL errors as the difference between these values and the ideal values, giving us

INL errors: $\{0 : -0.002 : -0.007 : -0.003 : -0.007 : -0.003 : 0.004 : 0\}$

For DNL errors, we find the difference between adjacent offset-free, gain-free, scaled values to give

DNL errors: $\{-0.002 : -0.005 : 0.004 : -0.004 : 0.004 : 0.007 : -0.004\}$

3. For absolute accuracy, we find the largest deviation between the measured values and ideal values, which, in this case, occurs at 0 V and is 11 mV. To relate this 11-mV value to effective bits, 11 mV should correspond to 1 LSB when $V_{\text{ref}} = 4$ V. In other words, we have the relationship

$$\frac{4 \text{ V}}{2^{N_{\text{eff}}}} = 11 \text{ mV} \tag{15.33}$$

which results in an absolute accuracy of $N_{\text{abs}} = 8.5$ bits.

4. For relative accuracy, we use the INL errors found in part 2, whose maximum magnitude is 0.007 LSB, or equivalently, 3.5 mV. We relate this 3.5-mV value to effective bits in the same manner as in part 3, resulting in a relative accuracy of $N_{\text{rel}} = 10.2$ bits.

EXAMPLE 15.5

A full-scale sinusoidal waveform is applied to a 12-bit A/D converter, and the output is digitally analyzed. If the fundamental has a normalized power of 1 W while the remaining power is 0.5 μW, what is the effective number of bits for the converter?

Solution

In this case, the signal-to-noise and distortion ratio is found to be

$$\text{SNDR} = 10 \log\left(\frac{1}{0.5 \times 10^{-6}}\right) = 63 \text{ dB} \tag{15.34}$$

Substituting this SNDR value into (15.29), we find

$$N_{eff} = \frac{63 - 1.76}{6.02} = 10.2 \text{ effective bits} \tag{15.35}$$

15.6 KEY POINTS

- In Nyquist-rate converters, there is a one-to-one correspondence between input and output values. Oversampling converters increase SNR by operating much faster than the input signal's Nyquist rate and filtering out quantization noise outside of the signal bandwidth. [p. 606]
- Data converters translate between digital binary codes and analog voltages or currents. In an ideal converter, the binary codes correspond to analog signal quantities that are precisely equally spaced. [p. 607]
- As long as it is not overloaded, the ideal quantizer introduces an additive quantization noise less than 1/2 of an LSB in magnitude. [p. 609]
- Quantization noise may be modeled as a random quantity uniformly distributed between –LSB/2 and +LSB/2. For a sinusoidal input, this results in a maximum SNR of 6.02N + 1.76 dB for an ideal N-bit converter. [p. 611]
- Integral nonlinearity (INL) is a popular measure of accuracy that specifies the deviation of a converter's input–output relationship away from the ideal, or least-squares fit, linear relationship. [p. 616]
- Both DNL and INL are measured at dc or very low input frequencies. They are therefore referred to as measures of static nonlinearity. [p. 616]
- Uncertainty in the sampling times of a converter, called aperture jitter, can be a significant performance limitation at high input signal frequencies. [p. 617]
- The most common measure of accuracy at high input signal frequencies is dynamic range, often quantified by the maximum achievable signal-to-noise and distortion ratio (SNDR). This may be expressed in dB, or as an effective number of bits. A converter's effective resolution bandwidth specifies the band over which SNDR is within 3 dB of its highest value. [p. 618]

15.7 REFERENCES

Analog Devices. *Analog-Digital Conversion Handbook.* Prentice Hall, Englewood Cliffs, New Jersey, 1986.

A. B. Grebene. *Bipolar and MOS Analog Integrated Circuit Design.* John Wiley & Sons, New York, 1984.

R. van de Plassche. *Integrated Analog-to-Digital and Digital-to-Analog Converters.* Kluwer Academic Publishers, Dordrecht, the Netherlands, 1994.

15.8 PROBLEMS

15.8.1 Section 15.1: Ideal D/A Converter

15.1 For an ideal 10-bit, unipolar D/A converter with $V_{LSB} = 1$ mV what is the largest output voltage?

15.8.2 Section 15.2: Ideal A/D Converter

15.2 What is the SNR for an ideal 12-bit unipolar A/D converter with $V_{ref} = 3$ V, when a sinusoidal input of 1 V_{pp} is applied? What size input would result in an SNR of 0 dB?

15.8.3 Section 15.3: Quantization Noise

15.3 An ideal quantizer is to cover the range from 0 to 1 V with less than 10 μV_{rms} quantization noise power. How many bits of resolution does the converter require?

15.4 A quantizer has an LSB step size of 3 mV and 10 bits of resolution. What is the maximum SQNR it can achieve?

15.5 An ideal quantizer takes an 0.8 V_{pp} sinusoid without overloading and quantizes it with an LSB step size of 5 mV. What is the resulting SQNR?

15.6 An input signal arrives at a quantizer already corrupted with some noise and having a SNR of 35 dB. How many bits of resolution does the quantizer require to ensure that the quantization noise is at least 3 dB smaller than the input noise?

15.8.4 Section 15.4: Signed Codes

15.7 Find the equivalent of (15.19) for 2's-complement coding.

15.8 Show that, by using 4-bit 2's-complement words, the correct final sum is obtained when the numbers +5, +5, and –7 are added together while any overflow effects are ignored.

15.9 Starting with two 4-bit 2's-complement words, we want to add +5 and +7 to obtain the correct answer of +12 with a 5-bit word. Show how an extra bit can be added at the left of each of the 4-bit words such that numbers up to ±15 can be represented. This approach is called *sign extension* and can be used to increase the word size of any number.

15.10 What is the representation of +8 and –8 in 5-bit 2's complement? Assuming a circuit added only one of these two numbers to an arbitrary 5-bit word, show a simple logic circuit (i.e., nand, nor, xnor, etc.) that would accomplish such an addition. (*Hint:* Note that the 4 LSBs in +8 and –8 are all 0 and thus do not need to be added.)

15.8.5 Section 15.5: Performance Limitations

15.11 The following measurements are found from a 3-bit unipolar D/A converter with $V_{ref} = 8$ V: (–0.01, 1.03, 2.02, 2.96, 3.95, 5.02, 6.00, 7.08). In units of LSBs, find the offset error, gain error, maximum DNL, and maximum INL.

15.12 How many bits of absolute accuracy does the converter in Problem 15.11 have? How many bits of relative accuracy does it have?

15.13 A 10-bit A/D converter has a reference voltage, V_{ref}, tuned to 10.24 V at 25 °C. Find the maximum allowable temperature coefficient in terms of $(\mu V)/°C$ for the reference voltage if the reference voltage is allowed to cause a maximum error of $(\pm 1/2)$ LSB over the temperature range of 0 to 50 °C.

15.14 Consider the following measured voltage values for a 2-bit D/A with a reference voltage of 4 V:

$$\{00 \leftrightarrow 0.01 \text{ V}\} \quad \{01 \leftrightarrow 1.02 \text{ V}\} \quad \{10 \leftrightarrow 1.97 \text{ V}\} \quad \{11 \leftrightarrow 3.02 \text{ V}\}$$

In units of LSB, find the offset error, gain error, worst absolute and relative accuracies, and worst differential nonlinearity. Restate the relative accuracy in terms of an N-bit accuracy.

15.15 Find the maximum magnitude of quantization error for a 12-bit A/D converter having V_{ref} equal to 5 V and 0.5-LSB absolute accuracy.

15.16 What sampling-time uncertainty can be tolerated for a 16-bit A/D converter operating on an input signal from 0–20 kHz?

15.17 The SNDR of an A/D converter is measured while varying the input signal and the results are shown in Fig. P15.17. What is the effective resolution of this converter, N_{eff}, at low frequencies? What is its effective resolution bandwidth?

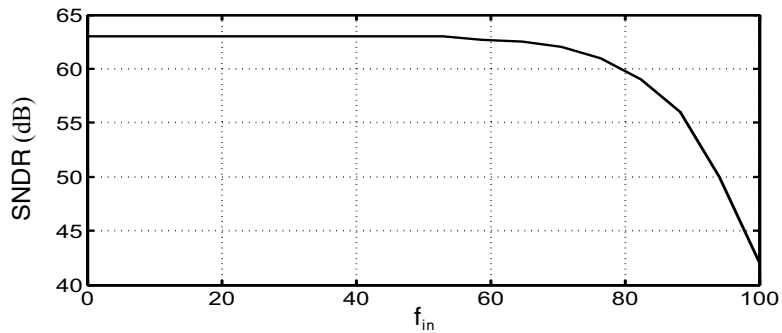


Fig. P15.17

CHAPTER

16 Nyquist-Rate D/A Converters

In this chapter, a variety of methods are discussed for realizing integrated Nyquist-rate digital-to-analog converters (DAC). Nyquist-rate D/A converters can be roughly categorized into four main types: decoder-based, binary-weighted, thermometer-code, and hybrid. Oversampling D/A converters are discussed separately in Chapter 18 due to their importance and because they are best described using many signal processing concepts.

16.1 DECODER-BASED CONVERTERS

Perhaps the most straightforward approach for realizing an N-bit D/A converter is to create 2^N reference signals and pass the appropriate signal to the output, depending on the digital input word. We refer to such D/A converters as *decoder-based converters.*

Key Point: *Decoder-based D/A converters operate by creating a set of analog reference levels and selecting one by turning on appropriate switches selected by an input digital code.*

16.1.1 Resistor-String Converters

One of the first integrated MOS 8-bit D/A converters was based on selecting one tap of a segmented resistor string by a switch network [Hamadé, 1978].[1] The switch network was connected in a tree-like decoder, as in the 3-bit D/A converter shown in Fig. 16.1. Notice that there will be one, and only one, low-impedance path between the resistor string and the input of the buffer, and that path is determined by the digital input word, B_{in}. In a CMOS implementation, transmission gates might be used rather than n-channel switches. However, when only n-channel pass transistors are used, the transistor-tree decoder can be laid out quite compactly since no contacts are required in the tree. Also, an n-channel-only approach is not much different in speed than a CMOS transmission-gate implementation; a transmission-gate approach has extra drain and source capacitance to ground, but this extra capacitance is offset by the reduced switch resistance, which is due to the parallel combination of p-channel and n-channel transistors. In addition, a transmission-gate implementation can operate closer to the positive voltage supply.

With a resistor-string approach, if we assume the buffer's offset voltage does not depend on its input voltage, the D/A converter has guaranteed monotonicity since any tap on the resistor string must have a lower voltage than its upper, neighbor tap. Also, the accuracy of this D/A depends on the matching precision of R in the resistor string. Although this resistor-matching precision depends on the type of resistors used, the use of polysilicon resistors can result in up to approximately 10 bits of accuracy.

The delay through the switch network is the major limitation on speed. However, in a multiplying D/A, the delay through the resistor string would also be a major source of delay since V_{ref} would become a second input signal. A useful technique for estimating the settling-time behavior in RC type circuits (i.e., circuits that have only real-axis poles) is the *zero-value time-constant* approach described in Section 4.2.4 [Sedra, 1991].

1. In fact, the resistor-string D/A was used to realize an A/D converter in the reference.

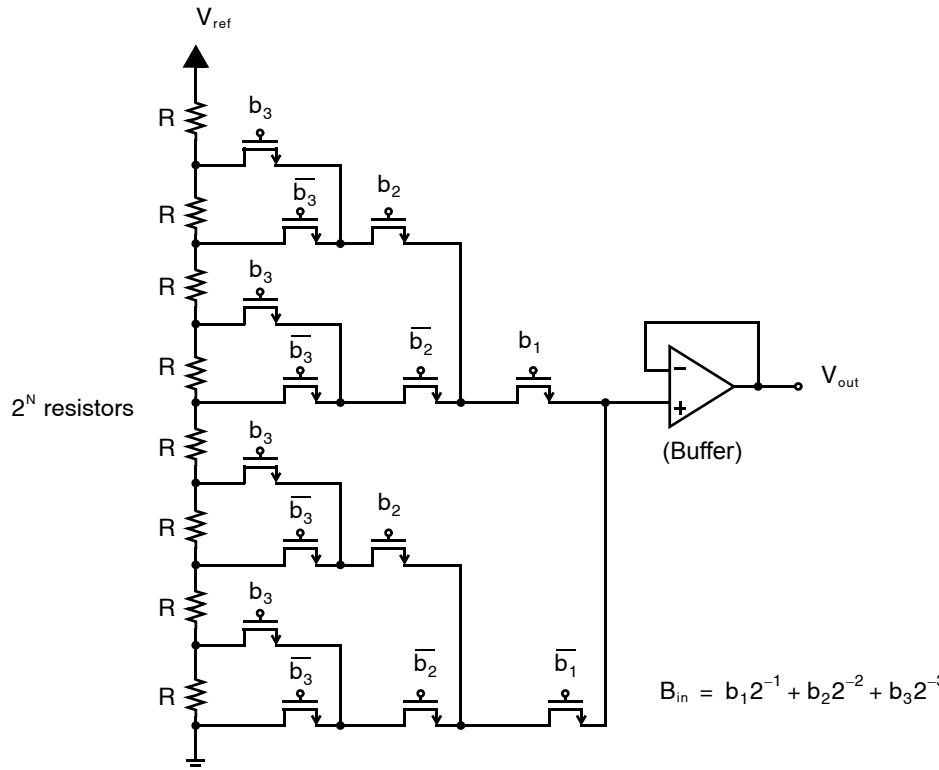


Fig. 16.1 Resistor-string 3-bit D/A converter with a transmission-gate, tree-like decoder.

Specifically, the dominant high-frequency time constant is estimated as the sum of the individual time constants due to each of the capacitances when all other capacitances are set to zero (i.e., replaced with open circuits). To find the individual time constant for a given capacitance, independent voltage sources are replaced with ground (independent current sources are opened), and the resistance seen by that capacitor is determined.

EXAMPLE 16.1

Show that an estimate of the time constant for a network of n resistors, each of size R in series, with capacitive loading C at each node (see Fig. 16.2), is given by $\tau = RC(n^2/2)$. How much settling time is required for the output to settle to 0.1 percent of its final value?

Solution

The zero-value time-constant due to the first capacitor on the left is simply RC. The second capacitor from the left has an individual time constant of $2RC$ and so on. Thus, the dominant high-frequency time constant, τ, is estimated as

$$\tau \cong RC(1 + 2 + \cdots + n)$$

The sum from 1 to n can be shown to be equal to $n(n+1)/2$ and, thus, for a large n, the dominant time constant

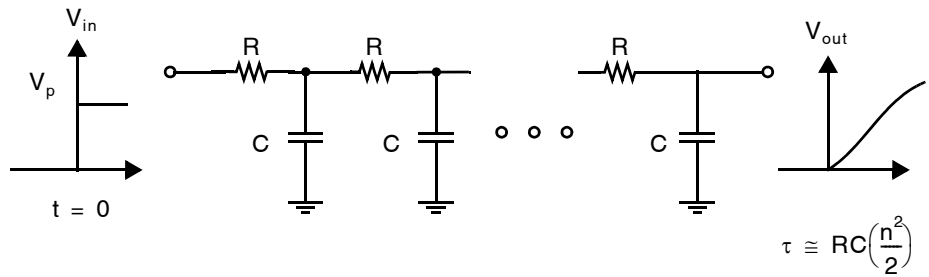


Fig. 16.2 Estimating the time constant for n resistors and capacitors in series.

can be estimated by

$$\tau \cong RC\left(\frac{n^2}{2}\right)$$

Using this time constant to estimate the output voltage charging behavior, we have

$$v_{out} \cong (1 - e^{-(t/\tau)})V_p$$

Thus, for v_{out} to equal $0.999\ V_p$, we find that a time of about 7τ is needed.

In a higher-speed implementation, logic can be used for the decoder, and a single bus is connected to a single resistor-string node, as shown in Fig. 16.3. This approach takes more area for the decoder and also results in a large capacitive loading on the single bus because the 2^N transistors' junctions are connected to the bus. However, if the digital decoder is pipelined, the D/A can be moderately fast.

16.1.2 Folded Resistor-String Converters

To reduce the amount of digital decoding and large capacitive loading, a *folded resistor-string D/A* can be used, as shown in Fig. 16.4 [Abrial, 1988]. This approach makes the decoding very similar to that for a digital memory, which reduces the total decoding area. To convert a digital input in this 4-bit example, the most significant bits, $\{b_1, b_2\}$, determine the single word line to be selected (all others will remain low). This operation connects a block of four adjacent resistor nodes to the four-bit lines. One of these bit lines is then connected to the output buffer by the bit-line decoder. Notice that the total number of transistor junctions on the output line is now $2\sqrt{2^N}$ because a set of transistors is connected directly to the output line and another set is connected to the chosen bit line. Thus, for a 10-bit converter, this approach would have a capacitive load of 64 junctions, as opposed to 1,024 junctions when we use the digital-decoding approach shown in Fig. 16.3. Unfortunately, the increase in speed is not equal to this large ratio since, when a word line goes high, all the bit lines must be pulled to new voltage levels—not just the one bit line connected to the output buffer.

16.1.3 Multiple Resistor-String Converters

In this variation, a second tapped resistor string is connected between buffers whose inputs are two adjacent nodes of the first resistor string, as shown in Fig. 16.5 [Holloway, 1984]. In the shown 6-bit example, the three

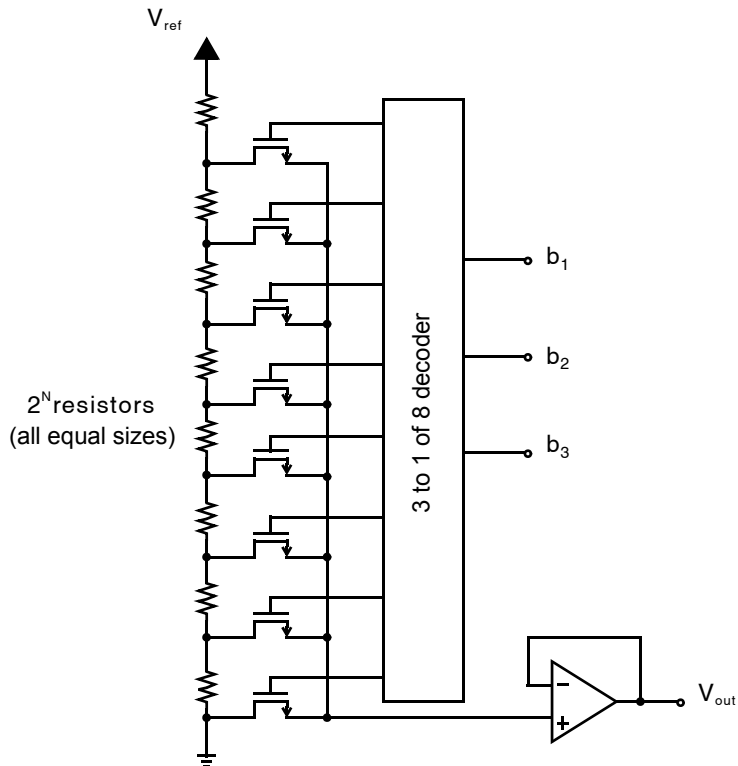


Fig. 16.3 Resistor-string 3-bit D/A converter with digital decoding.

MSBs determine which two adjacent nodes of the first resistor string are connected to the two intermediate buffers. The second resistor string linearly interpolates between the two adjacent voltages from the first resistor string. Finally, the output is determined by the lower LSBs where extra logic must take into account that sometimes the top intermediate buffer has the higher voltage, whereas at other times it has the lower voltage. This approach requires only $2 \times 2^{N/2}$ resistors, making it suitable for higher-resolution, low-power applications. This approach also has guaranteed monotonicity, assuming the opamps have matched, voltage-insensitive offset voltages. However, the opamps must be fast and low noise, which can be achieved using a BiCMOS process. Another point to note here is that, since the second resistor string is used to decode only the lower-order bits, the matching requirements of the second resistor string are not nearly as severe as those for the first string.

EXAMPLE 16.2

Assume that the first resistor string of a 10-bit, multiple-R-string, D/A converter must match to 0.1 percent, and that the first string realizes the top 4 bits. What is the matching requirement of the second resistor string, which realizes the lower 6 bits?

Solution

Errors in the first resistor string correspond directly to errors in the overall D/A output. However, since the second resistor string forms the lower LSB bits (6 bits, in this case), errors in the matching of these resistors

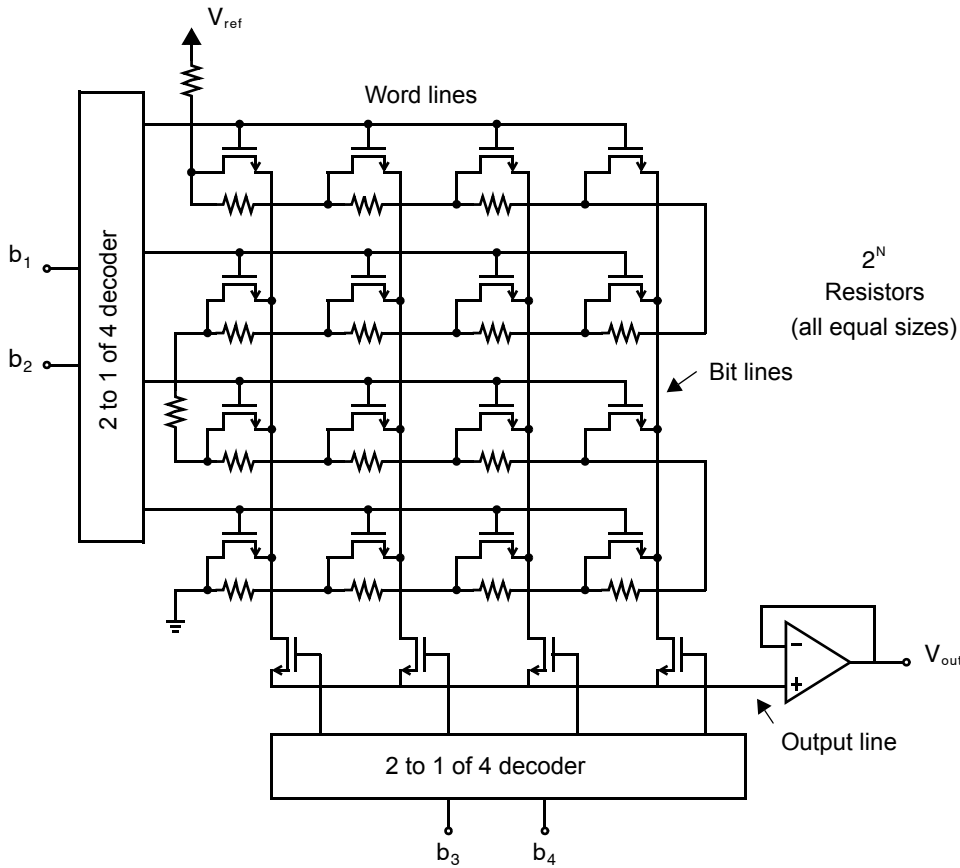


Fig. 16.4 A 4-bit folded resistor-string D/A converter.

cause an error only in the LSB portion of the output voltage. As a result, the second resistor string need match only to

$$2^4 \times 0.1\% = 1.6\% \tag{16.1}$$

16.1.4 Signed Outputs

In applications where negative output voltages are required, the bottom of the resistor string can be connected to $-V_{ref}$. This requires a negative power supply, and the circuit needed to realize a dual power supply with exactly matched voltages is nontrivial since any error in matching will result in an offset error. Many papers on D/A converters assume that $-V_{ref}$ is available but do not explain how it was obtained. If it is obtained off chip, the cost is significantly higher. One possibility is to use a switched-capacitor gain amplifier, as shown in Fig. 16.6, where a negative output can be realized by changing the clock phases of the input switches so an inverting amplifier is realized [Martin, 1987]. Another possibility is to sense the center tap of the resistor string and adjust $-V_{ref}$ to get the center tap voltage equal to zero volts.

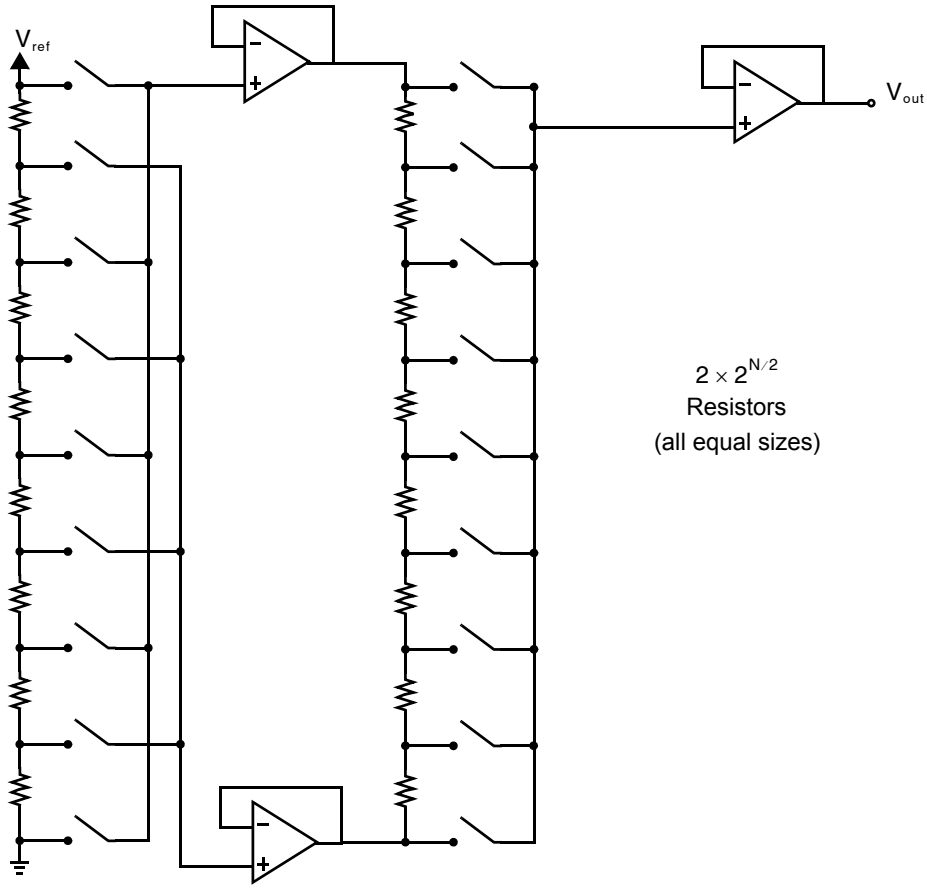


Fig. 16.5 Multiple R-string 6-bit D/A converter.

16.2 BINARY-SCALED CONVERTERS

Key Point: *Binary-scaled D/A converters combine binary-weighted circuit quantities (currents, resistors, capacitors, etc.) under digital control to realize an analog quantity. Such techniques are generally hardware-efficient, but can be subject to significant nonlinearities.*

The most popular approach for realizing at least some portion of D/A converters is to combine an appropriate set of signals that are all related in a binary fashion. This binary array of signals might be currents (in resistor or current approaches), but binary-weighted arrays of charge are also commonly used. In this section, resistor approaches are first discussed, followed by charge redistribution and current mode. Monotonicity is not guaranteed under such schemes since totally disparate sets of components are used to translate neighboring digital code words into analog voltages or currents, and large DNL is generally observed when large portions of the binary-weighted array are switched on and off. For example, this mismatch effect is usually the largest in a binary-array converter when the MSB is changed, resulting in the largest DNL at this location.

16.2.1 Binary-Weighted Resistor Converters

Binary-weighted resistor converters are popular for a bipolar technology so that bipolar differential pairs can be used for current switches. The basic architecture for a 4-bit converter is shown in Fig. 16.7.

If b_i is a 1, then the current to the ith resistor comes from the virtual ground of the opamp; otherwise, it comes from ground. Therefore, we have

$$\begin{aligned} V_{out} &= -R_F V_{ref}\left(-\frac{b_1}{2R} - \frac{b_2}{4R} - \frac{b_3}{8R} - \cdots\right) \\ &= \left(\frac{R_F}{R} V_{ref}\right) B_{in} \end{aligned} \tag{16.2}$$

where

$$B_{in} = b_1 2^{-1} + b_2 2^{-2} + b_3 2^{-3} + \cdots \tag{16.3}$$

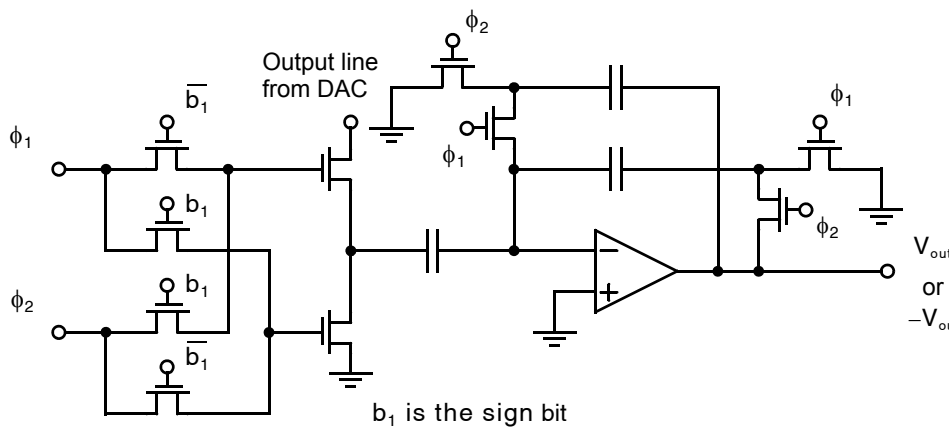


Fig. 16.6 Using an SC gain amplifier to realize a signed output from a unipolar (positive) D/A converter output: b_1 high causes a negative output.

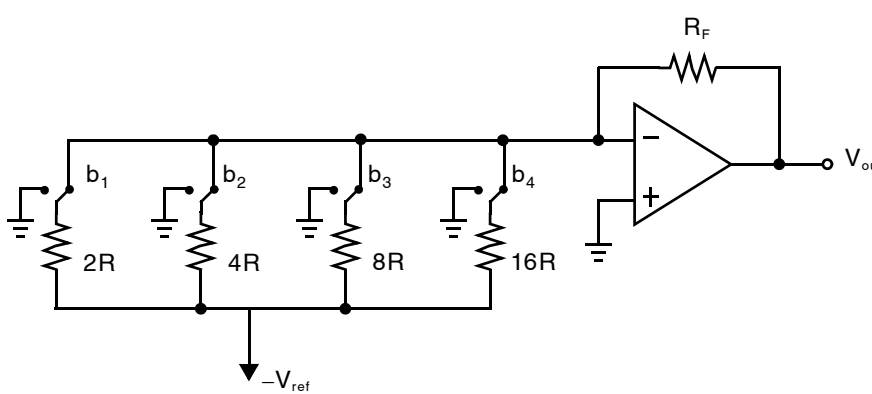


Fig. 16.7 Binary-weighted 4-bit resistor D/A converter.

Although this approach does not require many resistors or switches, it does have some disadvantages. The resistor and current ratios are on the order of 2^N, which may be large, depending on N. This large current ratio requires that the switches also be scaled so that equal voltage drops appear across them for widely varying current levels. Also, monotonicity is not guaranteed. Finally, this approach is prone to glitches for high-speed operation, as discussed on page 633.

16.2.2 Reduced-Resistance-Ratio Ladders

To reduce the large resistor ratios in a binary-weighted array, signals in portions of the array can be scaled by introducing a series resistor, as shown in Fig. 16.8. Here, note that the voltage node, V_A, is equal to one-fourth the reference voltage, V_{ref}, as a result of inserting the series resistor of value $3R$. Also note that an additional $4R$ resistor was added (to ground) such that resistance seen to the right of the $3R$ resistor equals R. Straightforward analysis shows that this converter has the same relationship to the binary digital signals as in the previous binary-weighted case, but with one-fourth the resistance ratio. Note, however, the current ratio has remained unchanged. Finally, note that, by repeating this procedure recursively to a binary-weighted ladder, one arrives at a structure commonly referred to as an *R-2R ladder*, described next.

16.2.3 R-2R-Based Converters

A very popular architecture for D/A converters uses R-2R ladders. These ladders are useful for realizing binary-weighted currents with a small number of components and with a resistance ratio of only 2, independent of the number of bits, N.

Consider the R-2R ladder network shown in Fig. 16.9. Analysis gives

$$\begin{aligned} R_4' &= 2R \\ R_4 &= 2R \parallel 2R = R \\ R_3' &= R + R_4 = 2R \\ R_3 &= 2R \parallel R_3' = R \end{aligned} \tag{16.4}$$

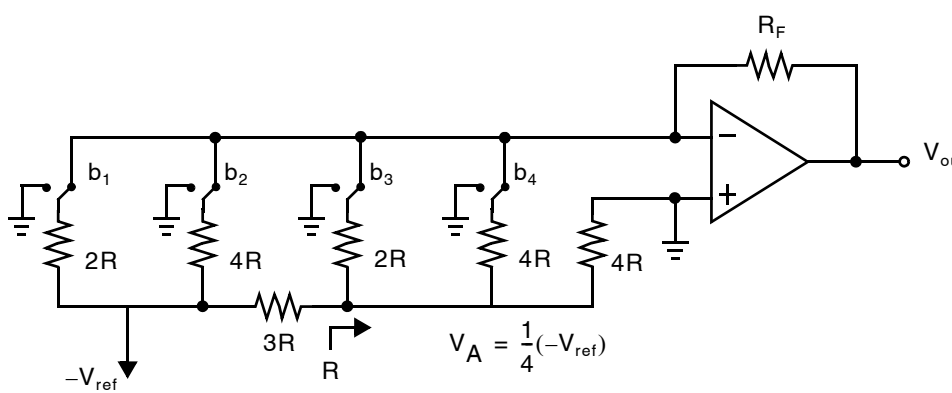


Fig. 16.8 Reduced-resistance-ratio 4-bit D/A converter.

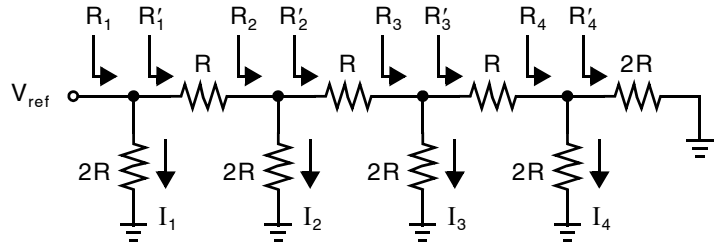


Fig. 16.9 R-2R resistance ladder.

and so on. Thus, $R_i' = 2R$ for all i. This result gives the following current relationships:

$$I_1 = \frac{V_{ref}}{2R} \tag{16.5}$$

Also, the voltage at node 2 is one-half the voltage at node 1, giving

$$I_2 = \frac{V_{ref}}{4R} \tag{16.6}$$

At node 3, the voltage divides in half once again, therefore

$$I_3 = \frac{V_{ref}}{8R} \tag{16.7}$$

and so on.

Thus, the R-2R ladder can be used to obtain binary-weighted currents while using only a single-size resistor. (The resistors of size 2R are made out of two resistors of size R, to improve matching properties.) As a result, this R-2R approach usually gives both a smaller size and a better accuracy than a binary-sized approach.

A 4-bit D/A converter that uses an R-2R ladder is shown in Fig. 16.10. For this R-2R-based circuit, we see that

Key Point: *R-2R ladder D/A converters produce binary-weighted currents without the need for a complete array of binary-weighted resistors. Instead, only 2:1 component ratios are needed, a significant savings for high-resolution converters.*

$$I_r = V_{ref}/(2R) \tag{16.8}$$

and

$$V_{out} = R_F \sum_{i=1}^{N} \frac{b_i I_r}{2^{i-1}} = V_{ref}\left(\frac{R_F}{R}\right) \sum_{i=1}^{N} \frac{b_i}{2^i} \tag{16.9}$$

However, as already mentioned, although the resistance ratio has been reduced, the current ratio through the switches is still large, and thus the switch sizes are usually scaled in size to accommodate the widely varying current levels. One approach to reduce this current ratio is shown in Fig. 16.11, where equal currents flow through all the switches. However, this configuration is typically slower since the internal nodes of the R-2R ladder now exhibit some voltage swings (as opposed to the configuration in Fig. 16.10 where internal nodes all remain at fixed voltages).

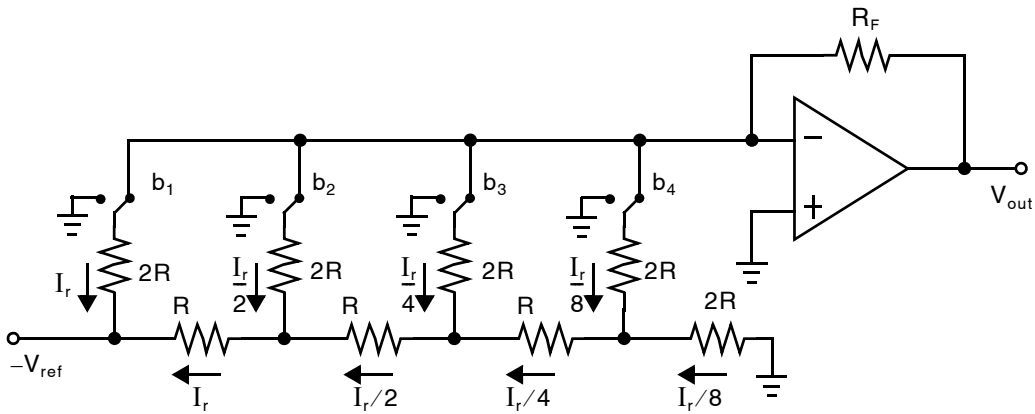


Fig. 16.10 4-bit R-2R based D/A converter.

16.2.4 Charge-Redistribution Switched-Capacitor Converters

Key Point: *Charge-redistribution switched-capacitor D/A converters sample a fixed reference voltage onto a binary-weighted capacitor array whose effective capacitance is determined by the input digital code. The sampled charge is then applied to a switched-capacitor gain stage to produce the analog output voltage.*

The basic idea here is to simply replace the input capacitor of an SC gain amplifier by a *programmable capacitor array* (PCA) of binary-weighted capacitors, as shown in Fig. 16.12. As in the SC gain amplifier, the shown circuit is insensitive to opamp input-offset voltage, 1/f noise, and finite-amplifier gain. Also, an additional sign bit can be realized by interchanging the clock phases (shown in parentheses) for the input switches.

It should be mentioned here that, as in the SC gain amplifier, carefully generated clock waveforms are required to minimize the voltage dependency of clock feed-through, and a deglitching capacitor should be used. Also, the digital codes should be changed only when the input side of the capacitors are connected to ground, and thus the switching time is dependent on the sign bit, which requires some extra digital complexity.

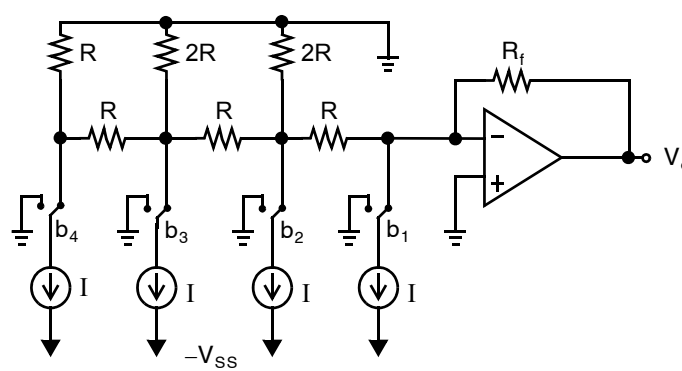


Fig. 16.11 R-2R ladder D/A converter with equal currents through the switches.

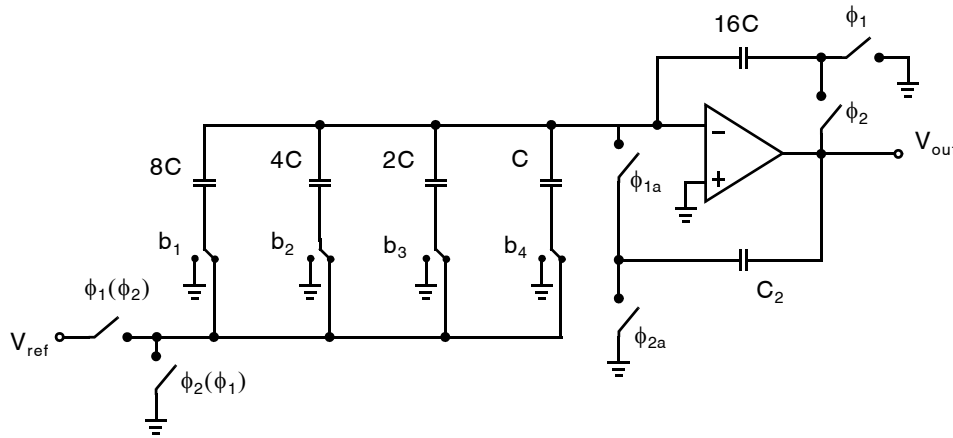


Fig. 16.12 Binary-array charge-redistribution D/A converter.

16.2.5 Current-Mode Converters

Current-mode D/A converters are very similar to resistor-based converters, but are intended for higher-speed applications. The basic idea is to switch currents to either the output or to ground, as shown in Fig. 16.13. Here, the output current is converted to a voltage through the use of R_F, and the upper portion of each current source always remains at ground potential. Implementations of current-mode D/A converters are more fully discussed in Section 16.3, where thermometer codes are introduced.

16.2.6 Glitches

Glitches are a major limitation during high-speed operation for converters that have digital logic, $\{b_1, b_2, \ldots, b_N\}$, directly related to switching signals. Glitches are mainly the result of different delays occurring when switching different signals. For example, when the digital input code changes from 0111 . . . 1 to 1000 . . . 0, all of the $N - 1$ LSBs turn off and the MSB turns on. However, it is possible that the currents due to the LSB switches will turn off slightly before the MSB current, causing the current to temporarily fall to zero. Alternatively, if the LSB switches

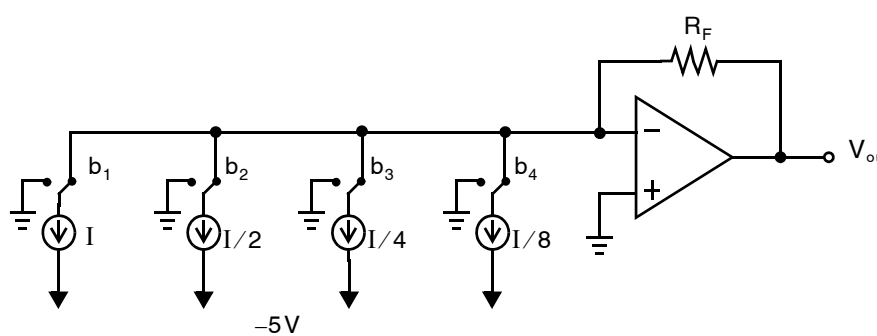


Fig. 16.13 Binary-weighted current-mode D/A converter.

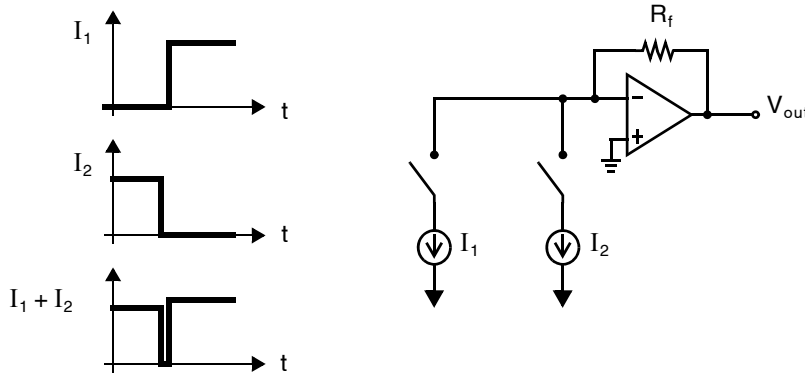


Fig. 16.14 Glitches. I_1 represents the MSB current, and I_2 represents the sum of the $N - 1$ LSB currents. Here, the MSB current turns off slightly early, causing a glitch of zero current.

turn off slightly after the MSB current, the current will temporarily go to its maximum value. In either case, a glitch occurs at the output unless these two delays are exactly matched, which is highly unlikely since the branches have different currents. This glitch phenomenon is illustrated in Fig. 16.14.

The glitch disturbance can be reduced by limiting the bandwidth (by placing a capacitor across the resistor R_f in Fig. 16.14), but this approach slows down the circuit. Another approach is to use a sample and hold on the output signal. Finally, the most popular way to reduce glitches is to modify some or all of the digital word from a binary code to a thermometer code, as discussed next.

16.3 THERMOMETER-CODE CONVERTERS

Key Point: *Thermometer-code converters employ an array of 2N-1 unitary equally-sized signal quantities. A digital thermometer code selects the appropriate number of these unitary signal quantities, summing them to produce the analog output with guaranteed monotonicity and better static linearity than binary-weighted converters.*

Another method for realizing a D/A converter is to digitally recode the input value to a thermometer-code equivalent. A *thermometer code* differs from a binary one in that a thermometer code has $2^N - 1$ digital inputs to represent 2^N different digital values. Clearly, a thermometer code is not a minimal representation since a binary code requires only N digital inputs to represent 2^N input values. However, as we will see, a thermometer-based converter does have advantages over its binary counterpart, such as low DNL errors, guaranteed monotonicity, and reduced glitching noise.

Typically, in a thermometer-code representation, the number of 1s represents the decimal value. For example, with a 3-bit binary input, the decimal value 4 is binary coded as 100 while its thermometer equivalent is 0001111. The codes for the remaining values in this 3-bit example are shown in Table 16.1.

One method to realize a D/A converter with the use of a thermometer-code input is to build $2^N - 1$ equal-sized resistors and switches attached to the virtual ground of an opamp, as shown in Fig. 16.15. Note that monotonicity is guaranteed here since, when the binary input changes to the next higher number, one more digital value in the thermometer code goes high, causing additional current to be drawn out of the virtual ground and forces the opamp output to go some amount higher (never lower). This is not necessarily the case for a binary-array D/A converter since mismatches between elements may cause the output to go lower even though the digital input value is increased.

Table 16.1 Thermometer-code representations for 3-bit binary values.

	Binary			Thermometer Code						
Decimal	b_1	b_2	b_3	d_1	d_2	d_3	d_4	d_5	d_6	d_7
0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	0	1	1
3	0	1	1	0	0	0	0	1	1	1
4	1	0	0	0	0	0	1	1	1	1
5	1	0	1	0	0	1	1	1	1	1
6	1	1	0	0	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1

Perhaps more importantly, a D/A converter based on a thermometer code greatly minimizes glitches, as compared to binary-array approaches, since banks of resistors are never exchanged at slightly different times when the output should change by only 1 LSB. It should also be mentioned here that latches can be used in the binary-to-thermometer code conversion such that no glitches occur in the digital thermometer-code words and pipelining can be also used to maintain a high throughput speed.

It is also of interest to note that the use of a thermometer code does not increase the size of the analog circuitry compared to a binary-weighted approach. In a 3-bit binary-weighted approach, the resistor values of R, $2R$, and $4R$ are needed for a total resistance of $7R$. This total value is the same as for the 3-bit thermometer-code approach shown in Fig. 16.15, and since resistors are created on an integrated circuit using area that is proportional to their size, each approach requires the same area (ignoring interconnect). The same argument can be used to show that the total area required by the transistor switches is the same since transistors are usually size-scaled in binary-weighted designs to account for the various current densities. All transistor switches in a thermometer-code approach are of equal sizes since they all pass equal currents. Finally, it should be mentioned that a thermometer-code charge-redistribution D/A can also be realized, as shown in Fig. 16.16.

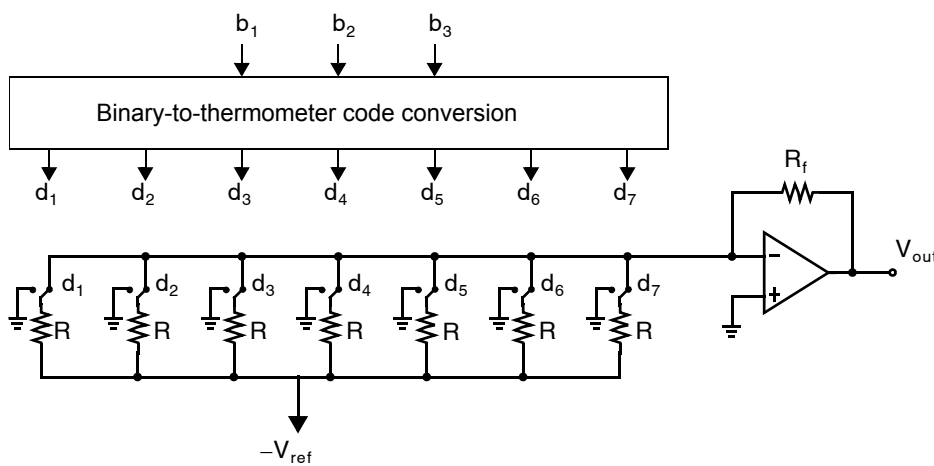


Fig. 16.15 A 3-bit thermometer-based D/A converter.

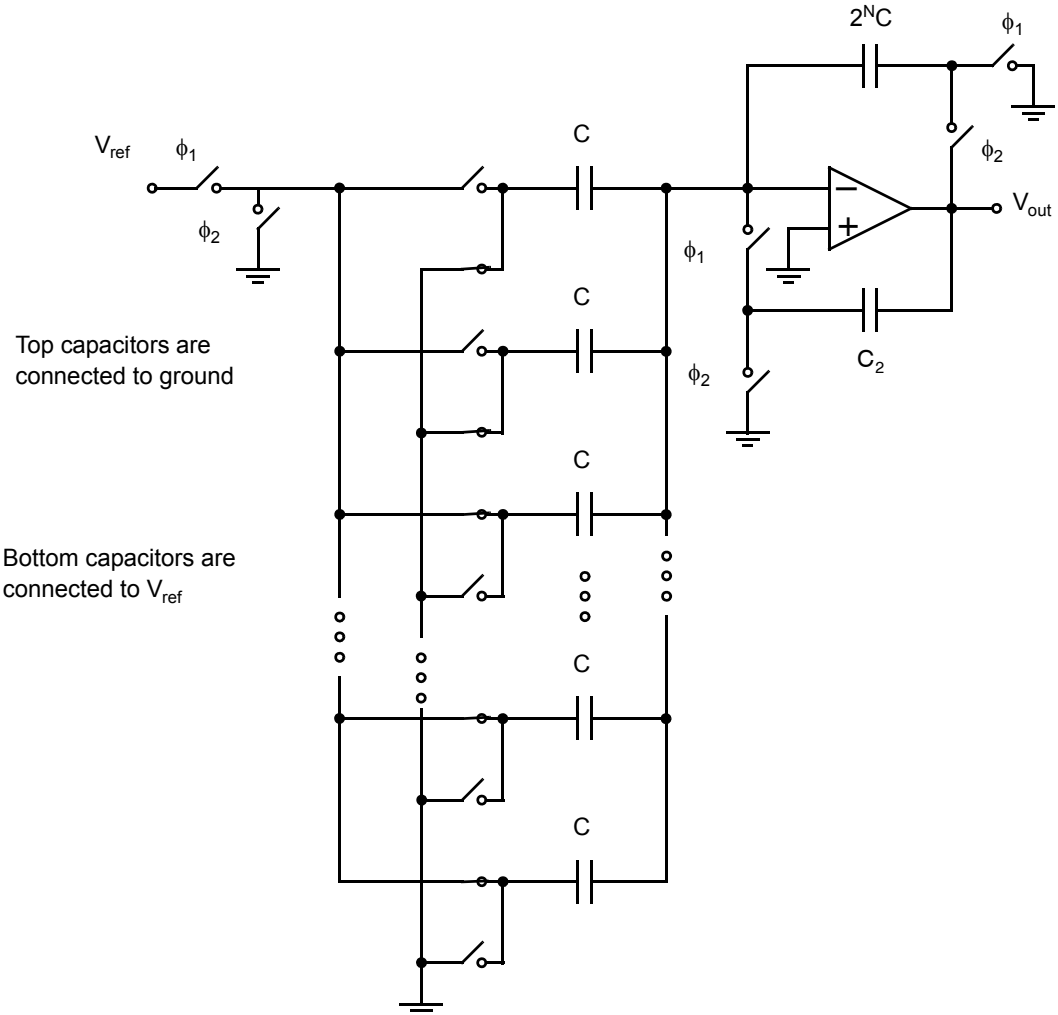


Fig. 16.16 Thermometer-code charge-redistribution D/A converter.

16.3.1 Thermometer-Code Current-Mode D/A Converters

A thermometer-code-based current-mode D/A converter, shown in Fig. 16.17, has been the basis for a variety of designs (see, for example, [Miki, 1986; Chi, 1986; and Letham, 1987]). Here, thermometer-code decoders are used for both the row and column decoders, resulting in inherent monotonicity and good DNL errors. Current is switched to the output when both row and column lines for a cell are high. Also, in high-speed applications, the output feeds directly into an off-chip 50-Ω or 75-Ω resistor, rather than an output opamp. Note that cascode current sources are used here to reduce current-source variation due to voltage changes in the output signal, V_{out}.

Note the need here for precisely timed edges of d_i and $\overline{d}_i$. If both logic levels are high simultaneously, V_{out} is shorted to ground. If both logic levels are low at the same time, the drain of Q_3 is pulled low and the circuit

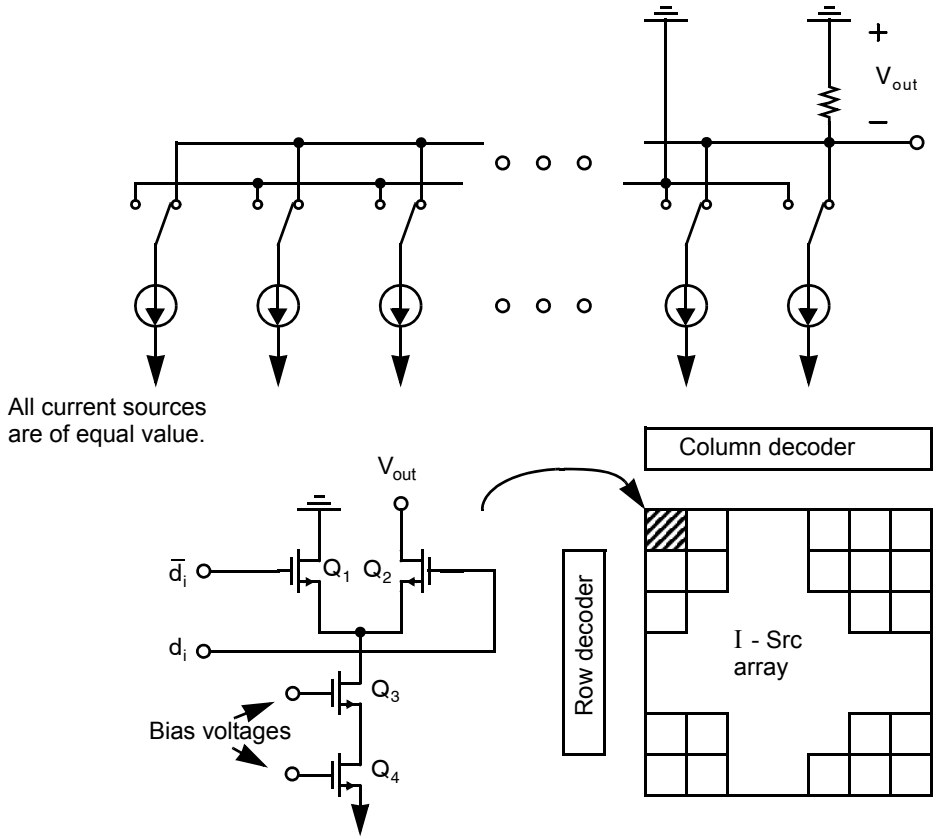


Fig. 16.17 Thermometer-code current-mode D/A converter. Q_1 and Q_2 form the current switch, whereas Q_3 and Q_4 implement a cascode current source.

takes longer to respond. To avoid the use of two logic driving levels, the gate of Q_2 should be connected to a dc bias voltage, as described in the next approach.

16.3.2 Single-Supply Positive-Output Converters

The architecture for a fast single-supply positive-output D/A (often used in video RAMs, which are known as RAMDACs) is shown in Fig. 16.18 [Colles, 1988]. Here, a matched feedback loop is used to set up accurate known current-source biasing. (Note that the opamp input connections appear reversed but are correct due to signal inversion by Q_4.) Also, to maintain accurate current matching that is independent of V_{out}, one side of each differential current-steering pair is connected to V_{bias}, rather than to the inversion of the bit signal. For example, when the current is steered to the output through Q_2, the drain-source voltage across the current source, Q_3, remains mostly constant if V_{out} stays near zero, such that Q_2 remains in the active region. Thus, Q_2 and Q_3 effectively form a *cascode current source* when they drive current to the output.

To maximize speed in this converter, the voltage swing at the common connections of the current switches (for example, Q_1, Q_2, and Q_3) should be small. To keep this swing small, this common connection should be at a voltage where the output transistors are just turned off when the current is steered to ground. It should also be noted that when

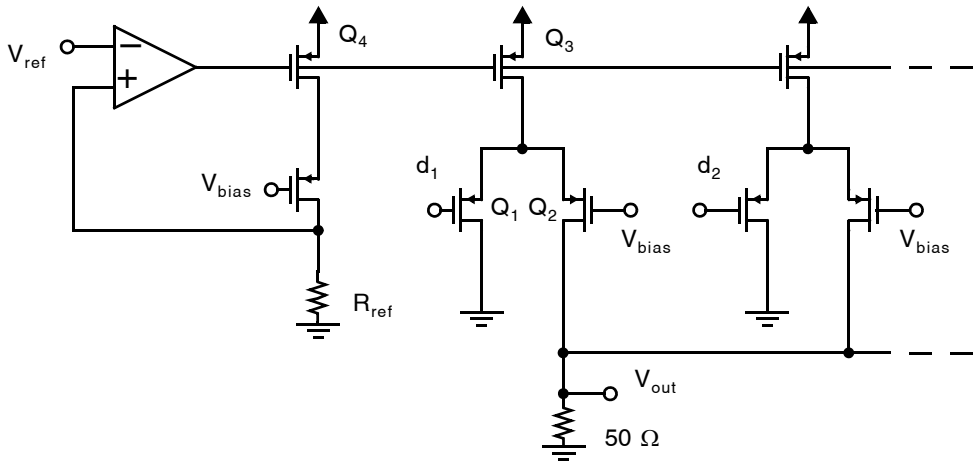


Fig. 16.18 A single-supply positive-output D/A converter.

a switch is turning on or off, the switching feedthrough from the digital input connected to the grounding transistor (for example, Q_1) actually enhances the switching. Finally, note that this design does not use two logic driving signals, d_i and $\overline{d}_i$, and can therefore be clocked at the maximum rate without the need for precisely timed edges.

16.3.3 Dynamically Matched Current Sources

The use of dynamic techniques with current switching is a method for realizing very well-matched current sources (up to 16-bit accuracy) for audio-frequency D/A converters [Schouwenaars, 1988].

This approach was used to design a 16-bit audio-frequency D/A converter, where the 6 MSB were realized using a thermometer code. Since the accuracy requirements are reduced for the remaining bits, a binary array was used in their implementation. The basic idea for realizing 63 accurately matched current sources for the 6 MSBs is illustrated in Fig. 16.19. Here, we want to set all the currents I_{di} to the same precise value, independent of transistor mismatches and charge injection. To accomplish this high degree of matching, each current source I_{di} is periodically calibrated with the use of a single reference current source, I_{ref}, through the use of the shift register. In other words, once calibration is accomplished on I_{d1}, the same current source, I_{ref}, is used to set the next current source, I_{d2}, to the same value as I_{d1}, and so on. Before proceeding to see how this calibration is accomplished, it is worth mentioning that the values of I_{di} need not precisely equal I_{ref} but do need to accurately match each other. Therefore, any common errors in the calibration stage are not a problem. Also, note that 64 current sources are calibrated, even though only 63 are required for the 6 MSB. This extra current source is needed so that the D/A converter can continuously operate, even when one of the current sources is being calibrated.

The method for calibrating and using one of the current sources is shown in Fig. 16.20, for I_{d1}. During calibration, the current source is connected to the reference current I_{ref}, while Q_1 is configured in a diode-connected mode. This places whatever voltage is necessary across the parasitic C_{gs} so that I_{d1} equals I_{ref}. When S_1 is opened, I_{d1} remains nearly equal to I_{ref}, assuming the drain-source voltage of Q_1 doesn't change and the clock feedthrough and charge injection of S_1 are small. During regular system use, the gate voltage (and therefore current) is determined by the voltage stored on the parasitic capacitance, C_{gs}.

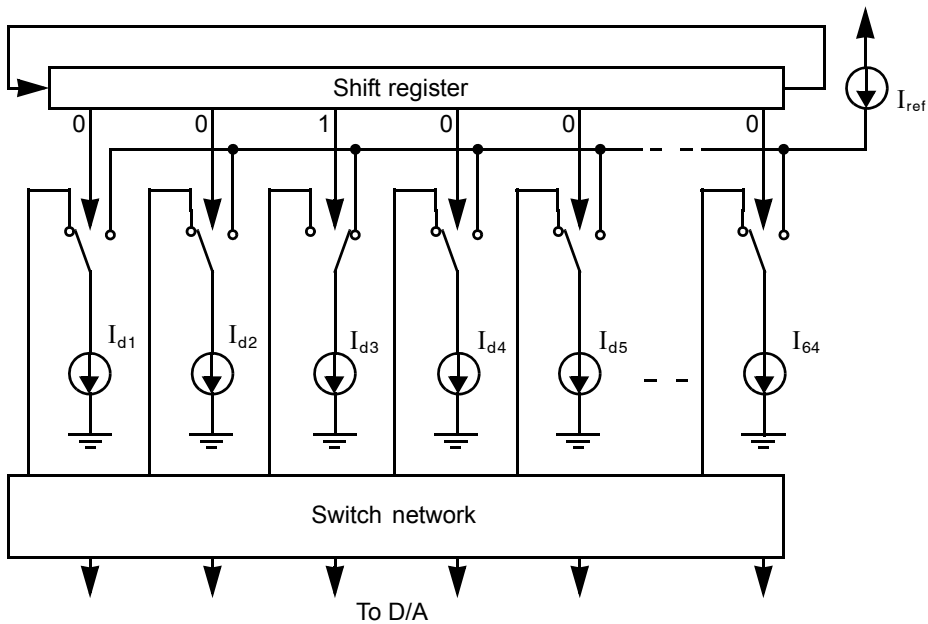


Fig. 16.19 Dynamically matching current sources for 6 MSB.

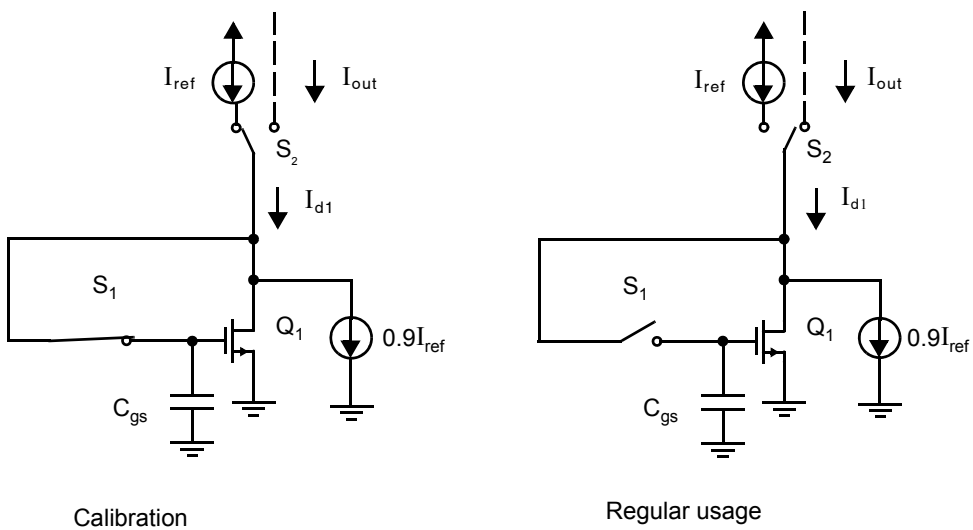


Fig. 16.20 Dynamically setting a current source, I_{d1}.

Key Point: *Selecting and summing current signals can provide high-speed D/A conversion. Output-impedance enhancement techniques may be needed to ensure the currents remain constat with respect to variations in output voltage. Calibration may also be used to ensure matching.*

A major limitation in matching the 64 current sources is due to the differences in clock feedthrough and charge injection of the switches S_i. Since mismatches will always exist between different switches, the best way to keep all current sources equal is to minimize the total amount of clock feedthrough and charge injection. These nonideal effects can be minimized by having the capacitance C_{gs} and bias voltage V_{GS} large. (A large V_{GS} voltage implies that any voltage difference will cause a smaller percentage current deviation.) To accomplish these requirements, the current source $0.9I_{ref}$ was added in parallel to Q_1, so that Q_1 needs only to source a current near $0.1I_{ref}$. With such an arrangement, a large, low-transconductance device can be used (a W/L = 1/8 might be used).

Finally, each current source must be recalibrated before the leakage currents (on the order of 10 pA/μm^2 of junction area) on C_{gs} cause the current source to deviate by 0.5LSB. Fortunately, having a large C_{GS} and V_{GS} has the added benefit of extending this calibration interval (every 1.7 ms in [Schouwenaars, 1988]).

Many other additional details are described in [Schouwenaars, 1988]. For example, during calibration, a p-channel-input common-gate amplifier is added to the diode-connected loop. This is typically done in dynamic switched-current circuits to control the drain-source voltage of Q_1 (i.e., to keep it constant independent of the actual current and to keep it matched to its value during regular use) and to speed up the circuit by decreasing the effect of parasitic capacitances on the large I_{ref} bus. Also, dummy switches are connected to S_1 to help minimize clock feedthrough by partially cancelling the charge injected.

16.4 HYBRID CONVERTERS

Key Point: *A hybrid converter combines different D/A architectures to realize different portions of the converter design.*

Combining the techniques discussed in Sections 16.1–16.3 for realizing different portions of a D/A converter results in hybrid designs. Hybrid designs are an extremely popular approach for designing converters because they combine the advantages of different approaches. For example, it is quite common to use a thermometer-code approach for the top few MSBs while using a binary-scaled technique for the lower LSBs. In this way, glitching is significantly reduced and accuracy is high for the MSB where it is needed most. However, in the LSBs where glitching and accuracy requirements are much reduced, valuable circuit area is saved with a binary-scaled approach. This section discusses some useful hybrid designs.

16.4.1 Resistor–Capacitor Hybrid Converters

It is possible to combine tapped resistor strings with switched-capacitor techniques in a number of different ways. In one approach, an SC binary-weighted D/A converter has its capacitors connected to adjacent nodes of a resistor-string D/A converter, as shown in Fig. 16.21 [Yang, 1989]. Here, the top 7 bits determine which pair of voltages across a single resistor is passed on to the 8-bit capacitor array. For example, if the top 7 bits are 0000001, then switches S_1 and S_2 would be closed, while the other S_i switches would remain open. The capacitor array then performs an 8-bit interpolation between the pair of voltages by connecting the capacitors associated with a 1 to the higher voltage and the capacitors associated with a 0 to the lower voltage. This approach gives guaranteed monotonicity, assuming the capacitor array is accurate to only 8 bits. The converter in [Yang, 1989] has 15-bit monotonicity, without trimming, and 100-kHz sampling frequencies for a very-low-power 10-mW realization.

16.4.2 Segmented Converters

The segmented D/A converter is a popular example of a hybrid converter [Schoeff, 1979; Grebene, 1984; Schouwenaars, 1988]. A 6-bit segmented D/A converter is shown in Fig. 16.22. In this approach, the two MSB currents are

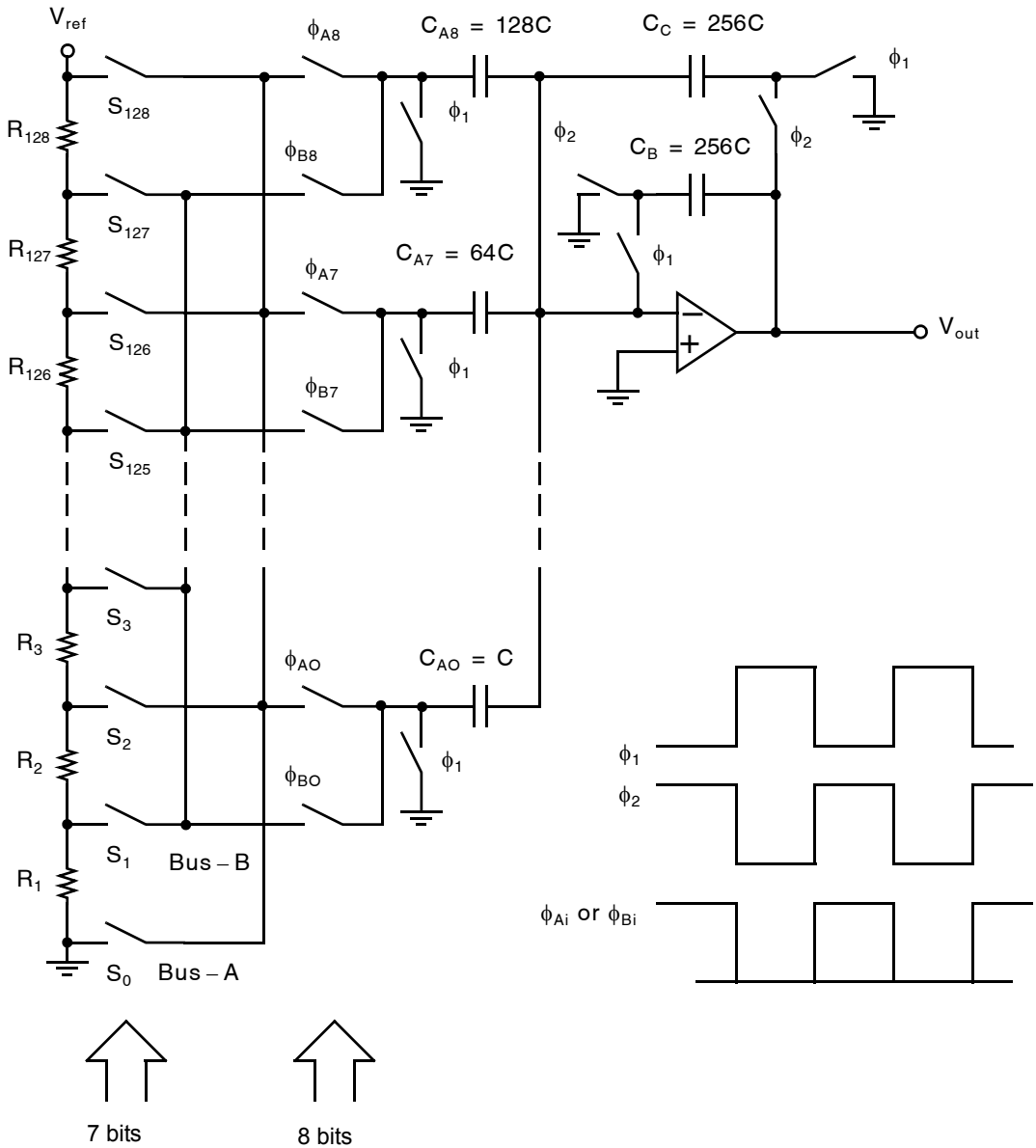


Fig. 16.21 One example of a 15-bit resistor–capacitor hybrid D/A converter.

obtained in one segment from three equal current sources using thermometer coding. High bits are switched to the output, whereas low bits are switched to ground. As discussed in Section 16.3, the use of a thermometer code for the MSB currents greatly minimizes glitches. For the four LSBs, one additional current source from the MSB segment is diverted where it is divided into binary-weighted currents, which are also switched to either ground or the output. Although this four-LSB segment is not guaranteed to be monotonic, its accuracy requirements are very relaxed, since it is used only for the LSBs.

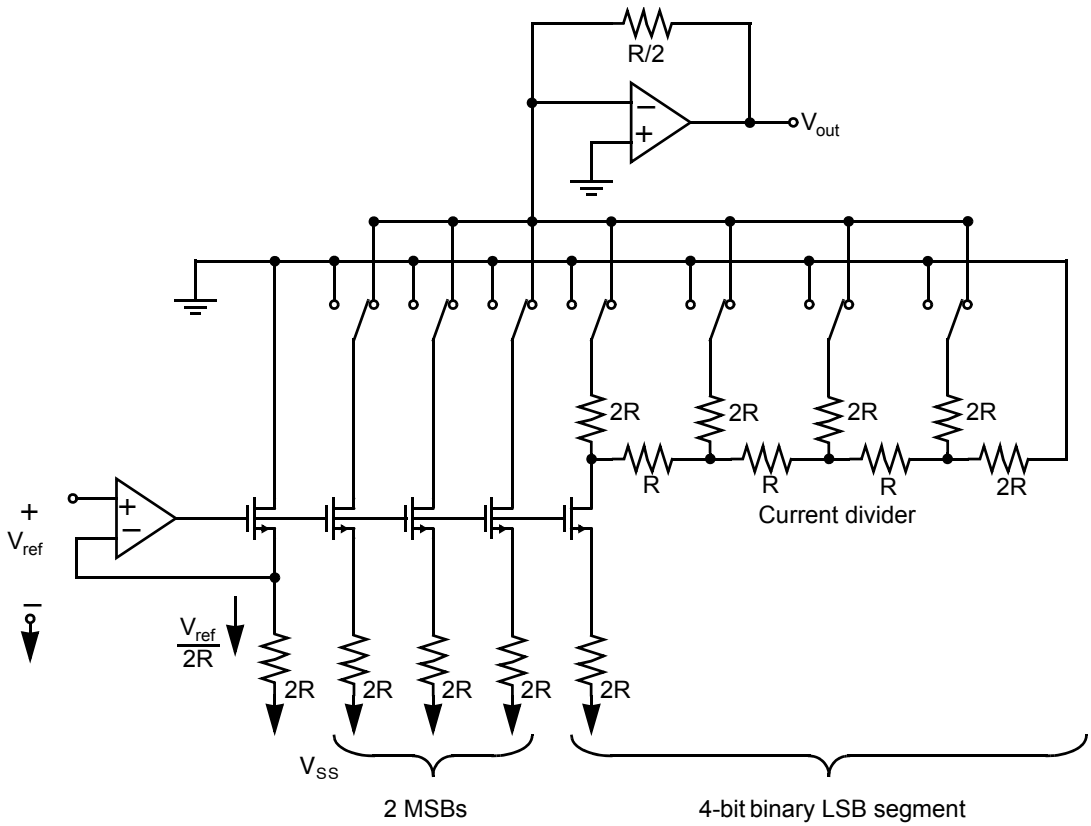


Fig. 16.22 A 6-bit segmented D/A converter.

16.5 KEY POINTS

- Decoder-based D/A converters operate by creating a set of analog reference levels and selecting one by turning on appropriate switches selected by an input digital code. [p. 623]
- Binary-scaled D/A converters combine binary-weighted circuit quantities (currents, resistors, capacitors, etc.) under digital control to realize an analog quantity. Such techniques are generally hardware-efficient, but can be subject to significant nonlinearities. [p. 628]
- R-2R ladder D/A converters produce binary-weighted currents without the need for a complete array of binary-weighted resistors. Instead, only 2:1 component ratios are needed, a significant savings for high-resolution converters. [p. 631]
- Charge-redistribution switched-capacitor D/A converters sample a fixed reference voltage onto a binary-weighted capacitor array whose effective capacitance is determined by the input digital code. The sampled charge is then applied to a switched-capacitor gain stage to produce the analog output voltage. [p. 632]
- Thermometer-code converters employ an array of $2N - 1$ unitary equally-sized signal quantities. A digital thermometer code selects the appropriate number of these unitary signal quantities, summing them to produce the analog output with guaranteed monotonicity and better static linearity than binary-weighted converters. [p. 634]

- Selecting and summing current signals can provide high-speed D/A conversion. Output-impedance enhancement techniques may be needed to ensure the currents remain constat with respect to variations in output voltage. Calibration may also be used to ensure matching. [p. 640]
- A hybrid converter combines different D/A architectures to realize different portions of the converter design. [p. 640]

16.6 REFERENCES

A. Abrial, et al. "A 27-MHz Digital-to-Analog Video Processor," *IEEE J. of Solid-S tate C ircuits*, Vol. 23, pp. 1358–1369, December 1988.

K. K. Chi, et al. "A CMOS Triple 100-Mbit/s Video D/A Converter with Shift Register and Color Map," *IEEE J. of Solid-State Circuits*, Vol. 21, pp. 989–996, December 1986.

J. Colles. "TTL to CMOS Voltage Level Translator," patent no. 4,794,282, granted December 27, 1988.

A. B. Grebene. *Bipolar and MOS Analog Integrated Circuit Design.* John Wiley & Sons, New York, 1984.

D. W. J. Groeneveld, H. J. Schouwenaars, H. A. H. Termeer, and C. A. A. Bastiaansen. "A Self-Calibration Technique for Monolithic High-Resolution D/A Converters," *IEEE J. of Solid-State Circuits*, Vol. 24, pp. 1517–1522, December 1989.

A. R. Hamadé. "A Single-Chip All-MOS 8-bit A/D Converter," *IEEE J. of Solid-State Circuits*, Vol. 13, pp. 785–791, December 1978.

P. Holloway. "A Trimless 16-bit Digital Potentiometer," *IEEE Intl. Solid-State Circuits Conf.*, pp. 66–67, February 1984.

L. Letham, et al. "A High-Performance CMOS 70-MHz Palette/DAC," *IEEE J. of Solid-S tate Circuits*, Vol. 22, pp. 1041–1047, December 1987.

K. W. Martin, L. Ozcolak, Y. S. Lee, and G. C. Temes. "A Differential Switched-Capacitor Amplifier," *IEEE J. of Solid-S tate Circuits*, Vol. 22, pp. 104–106, February 1987.

T. Miki, et al. "An 80-MHz 8-bit CMOS D/A Converter," *IEEE J. of Solid-State Circuits*, Vol. 21, pp. 983–988, December 1986.

P. H. Saul and J. S. Urquhart. "Techniques and Technology for High-Speed D/A Conversion," *IEEE J. of Solid-S tate Circuits*, Vol. 19, pp. 62–68, December 1984.

A. S. Sedra and K. C. Smith. *Microelectronic Circuits*, 3rd ed. Saunders College Publishing/HRW, New York, 1991.

J. A. Schoeff. "An Inherently Monotonic 12-bit DAC," *IEEE J. of Solid-State Circuits*, Vol. 14, pp. 904–911, December 1979.

H. J. Schouwenaars, D. W. J. Groeneveld, and H. A. H. Termeer. "A Low-Power Stereo 16-bit CMOS D/A Converter for Digital Audio," *IEEE J. of Solid-State Circuits*, Vol. 23, pp. 1290–1297, December 1988.

J. W. Yang and K. W. Martin. "High-Resolution Low-Power CMOS D/A Converter," *IEEE J. of Solid-S tate Circuits*, Vol. 24, pp. 1458–1461, October 1989.

16.7 PROBLEMS

16.7.1 Section 16.1: Decoder-Based Converters

16.1 Derive an expression for the number of switches in a general N-bit resistor-string D/A converter similar to that shown in Fig. 16.1.

16.2 Assume that an 8-bit resistor-string D/A converter with digital decoding (see Fig. 16.3) has a total resistor-string resistance of 400 Ω, that its pass transistors have an on resistance of 400 Ω, and that the drain-source capacitances to ground of its pass transistors are 0.1 pF. Ignoring all other effects and using the zero-value time-constant approach, estimate the worst-case settling time to 0.1 percent.

16.3 Derive an expression for the number of switches in an N-bit folded resistor-string D/A converter similar to that shown in Fig. 16.4 (Assume N is even).

16.4 Assume that the first resistor string of a 10-bit multiple-R-string D/A converter must match to 0.1 percent whereas the second string must match to 1.6 percent since the converters realize the top 4 bits and lower 6 bits, respectively. For $V_{ref} = 5$ V, how much offsets in the opamps can be tolerated?

16.7.2 Section 16.2: Binary-Scaled Converters

16.5 For a binary-weighted 10-bit resistor D/A converter (see Fig. 16.7), assume that R_F is chosen such that the output goes from 0 to $V_{ref} - V_{LSB}$. What is the resistor ratio between the largest and smallest resistors? What is the ratio between the currents through the switches for b_1 and b_{10}?

16.6 It is desired to realize a binary-weighted 4-bit resistor D/A converter, as shown in Fig. 16.7, that must be linear to 10 bits. Ignoring all other nonidealities except for resistance mismatch error, what percentage matching accuracy is required for each of the $b_2 - b_4$ resistors relative to the b_1 resistor?

16.7 Consider an 8-bit D/A converter built using binary-weighted capacitors, where capacitor tolerances are ± 0.5 percent. What would be the worst-case differential nonlinearity in units of LSBs, and at what transition does it occur?

16.8 Consider the reduced-resistance-ratio approach shown in Fig. 16.8, where a single series resistor is applied to an N-bit converter, and where N is an even number. What is the improvement in the resistance ratio compared to using all binary-scaled resistances?

16.9 Draw the circuit for a reduced-resistance-ratio 8-bit D/A converter (see Fig. 16.8), where a resistor is inserted between b_4 and b_5. Ignoring R_F, what is the resistance spread? Now draw a similar circuit, but add resistors between b_2 and b_3 as well as between b_6 and b_7. What is this circuit's resistance spread?

16.10 Assuming all the switches in Fig. P16.10 are MOSFETS and are scaled so that each has 100 mV across the drain source when on (including s_4, which always remains on), show that no accuracy is lost.

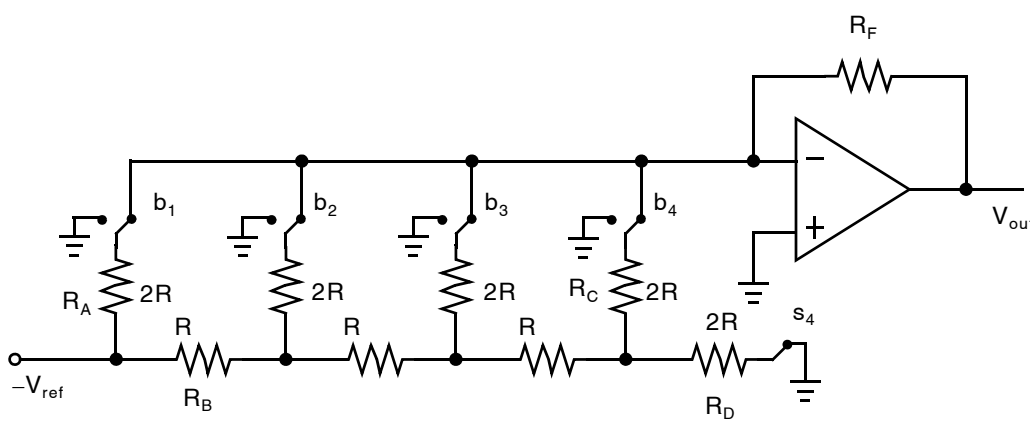


Fig. P16.10

16.11 For the 4-bit R-2R-ladder D/A converter shown in Fig. P16.10, what is the output error (in LSBs) when $R_A = 2.01R_B$? What is the output error when $R_C = 2.01R$?

16.12 Show that the D/A converter circuit shown in Fig. 16.11 operates correctly. Estimate the speed of this circuit if the opamp has infinite bandwidth, $R = 10$ kΩ, and all nodes have a capacitance of 0.5 pF to ground.

16.13 Consider the D/A converter shown in Fig. 16.11, where $I = 1$ mA, $R_f = 2$ kΩ and $R = 10$ kΩ. If this converter is perfectly linear but is found to have an offset error of 0.15 LSB and a gain error of 0.2 LSB, find the output levels for the inputs $b_1b_2b_3b_4$ equal to 0000, 1000, and 1111.

16.14 A 12-bit binary-weighted D/A converter produces a glitch voltage of 0.5 V when the MSB is changed. What is the glitch voltage reduced to if a thermometer-code approach is used for the top 4 bits, whereas binary-weighting is used for the remaining 8 bits?

16.7.3 Section 16.3: Thermometer-Code Converters

16.15 For the circuit in Fig. 16.17, sketch a typical waveform that would occur on the node connecting Q_1, Q_2, and Q_3 when the code signals, d_i and $\overline{d}_i$, are nonoverlapping. Repeat this sketch if the code signals are overlapping. Assume the code signals have logic swings from 0 to –5 V and that the threshold voltage for the transistors equals 1 V.

16.16 For the circuit in Fig. 16.18, sketch a typical waveform that would occur on the node connecting Q_1, Q_2, and Q_3 when the code signal, d_i, goes low-to-high and high-to-low. Assume here that the code signal has a logic swing of 0 to 5 V, that the threshold voltage for the transistors equals –1 V, and that V_{bias} is 1 V.

16.17 A D/A converter is realized using dynamically matched current sources, as shown in Fig. 16.20. Assuming all the transistors are ideal, find W/L for the Q_1 needed to set $V_{GS} = 3$ V when $I_{ref} = 50\ \mu A$, $V_t = 1$ V, and $\mu_n C_{ox} = 92\ \mu A/V^2$. If switch S_1 causes a random charge injection voltage of 1 mV, what is the expected percentage of random variation of the current being held on Q_1?

16.18 Repeat Problem 16.17 if the design does not incorporate the $0.9I_{ref}$ extra current source (i.e., if Q_1 must be the source for all of I_{ref}).

16.7.4 Section 16.4: Hybrid Converters

16.19 What is the maximum peak-to-peak output swing of the D/A converter in Fig. 16.22 in terms of V_{ref}?

16.20 What should the relative size (W/L) of the switches in Fig. 16.22 be in order to ensure they all carry the same current density?

CHAPTER

17 Nyquist-Rate A/D Converters

Architectures for realizing analog-to-digital (A/D) converters can be roughly divided into three categories (Table 17.1)—low-to-medium speed, medium speed, and high speed. In this chapter, design details are discussed for these different approaches except for oversampling converters. Oversampling converters are best described using many signal processing concepts and are therefore discussed separately in Chapter 18.

Before proceeding, it should be noted that when discussing the design of A/D converters, we usually ignore the 0.5 LSB offset present in the A/D transfer characteristic. Such a simplification is made so as not to complicate the concepts presented. Many of the converter architectures described make extensive use of switched capacitor circuits and comparators; the reader is referred to those chapters for more detail on those circuits.

17.1 INTEGRATING CONVERTERS

Integrating A/D converters is a popular approach for realizing high-accuracy data conversion on very slow-moving signals. These types of converters have very low offset and gain errors in addition to being highly linear. A further advantage of integrating converters is the small amount of circuitry required in their implementation. One application that has traditionally made use of integrating converters is measurement instruments such as voltage or current meters.

A simplified diagram for a *dual-slope* integrating converter is shown in Fig. 17.1. Dual-slope refers to this converter performing its conversion in two phases, (I) and (II), in the following manner (Fig. 17.2):

Phase (I) Phase (I) is a fixed time interval of length T_1 determined by running the counter for 2^N clock cycles. Thus, we have

$$T_1 = 2^N T_{clk} \tag{17.1}$$

where T_{clk} is the period for one clock cycle. During this interval, switch S_1 is connected to $-V_{in}$ such that V_x ramps up proportional to the magnitude of V_{in}. Assuming V_x is initially equal to zero (due to a pulse on S_2), and

Table 17.1 Different A/D converter architectures.

Low-to-Medium Speed, High Accuracy	Medium Speed, Medium Accuracy	High Speed, Low-to-Medium Accuracy
Integrating	Successive approximation	Flash
Oversampling	Algorithmic	Two-step
		Interpolating
		Folding
		Pipelined
		Time-interleaved

that V_{in} is constant, we have the following relationship for V_x:

$$V_x(t) = -\int_0^t \frac{(-V_{in})}{R_1C_1}\,d\tau = \frac{V_{in}}{R_1C_1}t \tag{17.2}$$

Thus, at the end of phase (I), the value of V_x is equal to $V_{in}T_1/R_1C_1$.

Phase (II) Phase (II) occurs for a variable amount of time, T_2, as shown in Fig. 17.2 for three different input voltages. At the beginning of this phase, the counter is reset and switch S_1 is connected to V_{ref}, resulting in a constant slope for the decaying voltage at V_x. To obtain the digital output value, the counter simply counts until V_x is less than zero, at which point that count value equals the digitized value of the input signal, V_{in}. Thus, assuming the digital output count is normalized so that the largest count is unity, the counter output, B_{out}, can be defined to be

$$B_{out} = b_1 2^{-1} + b_2 2^{-2} + \cdots + b_{N-1}2^{-(N-1)} + b_N 2^{-N} \tag{17.3}$$

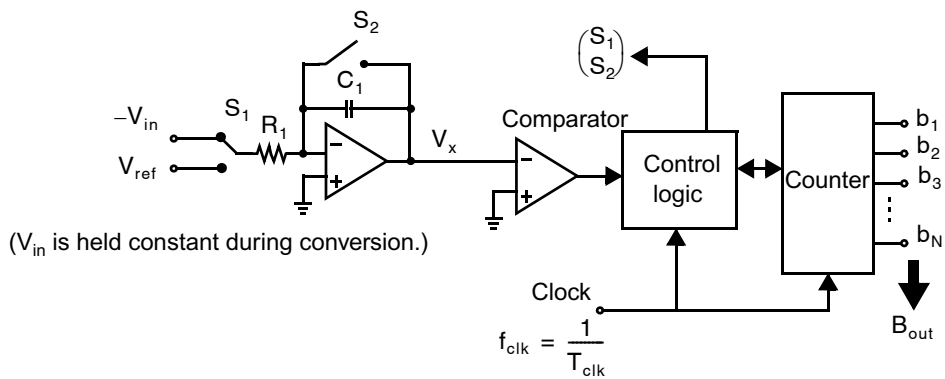


Fig. 17.1 Integrating (dual slope) A/D converter.

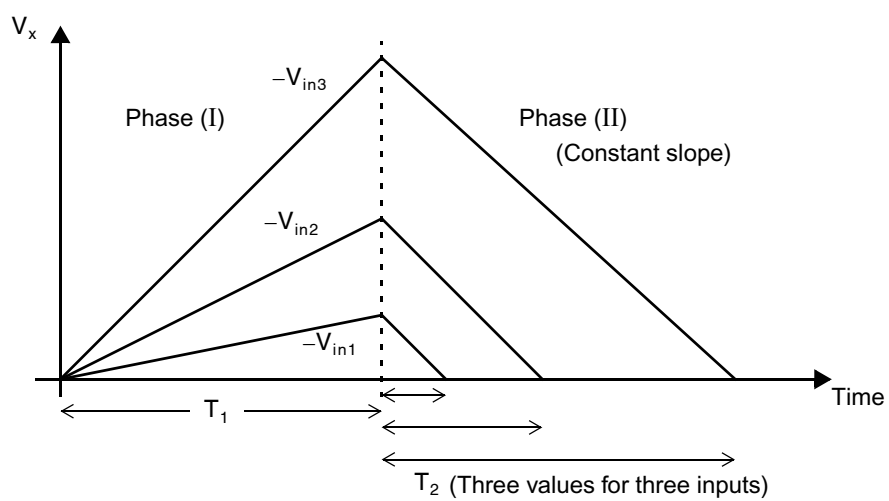


Fig. 17.2 Operation of the integrating converter for three different input voltages.

and we have

$$T_2 = 2^N B_{out} T_{clk} = (b_1 2^{N-1} + b_2 2^{N-2} + \cdots + b_{N-1} 2 + b_N) T_{clk} \tag{17.4}$$

To see why this count gives the correct value, we find the equation for V_x during phase (II) to be given by

$$\begin{aligned} V_x(t) &= -\int_{T_1}^{t} \frac{V_{ref}}{R_1 C_1} d\tau + V_x(T_1) \\ &= \frac{-V_{ref}}{R_1 C_1}(t - T_1) + \frac{V_{in} T_1}{R_1 C_1} \end{aligned} \tag{17.5}$$

Since V_x equals zero when $t = T_1 + T_2$, we can write

$$0 = \frac{-V_{ref} T_2}{R_1 C_1} + \frac{V_{in} T_1}{R_1 C_1} \tag{17.6}$$

and thus T_2 is related to T_1 by the following relationship:

$$T_2 = T_1 \left(\frac{V_{in}}{V_{ref}} \right) \tag{17.7}$$

Combining (17.7) with (17.1) and (17.4), we find

$$B_{out} = b_1 2^{-1} + b_2 2^{-2} + \cdots + b_{N-1} 2^{-(N-1)} + b_N 2^{-N} = \frac{V_{in}}{V_{ref}} \tag{17.8}$$

as expected.

Key Point: *Dual-slope integrating conversion proceeds first by integrating the input for a fixed time period, then applying a known input to the integrator and measuring the time required for the integrator output to return to zero.*

From (17.8), we see that in a dual-slope conversion (i.e., two phases), the digital output does not depend on the time constant, $R_1 C_1$. In fact, the value of this time constant need only be stable during a single conversion for proper operation. However, R_1 and C_1 should be chosen such that a reasonable large peak value of V_x is obtained without clipping to reduce noise effects. It is also possible to perform a single-slope conversion where only one integration phase is needed, however the integration time would then be a function of the time-constant value, $R_1 C_1$, and a gain error would most likely occur.

Although a dual-slope converter does not suffer from gain error, it can have an offset error due to opamp offset and other factors. Such an offset error can be calibrated out by going to a quad-slope conversion. In a quad-slope conversion, a dual-slope conversion is performed twice: once with the input connected to ground (or other known dc quantity), and then with the input connected to the signal to be converted, V_{in}. A subtraction of the two output words causes the offset error to be reduced to zero.

Key Point: *Dual-slope conversion does not require accuracy in defining the integration time constant. Offset errors can be calibrated by performing an additional conversion using a known dc input quantity.*

The conversion speed for these types of converters is quite slow. For example, in a dual-slope converter, the worst-case conversion speed occurs when V_{in} equals V_{ref}. In this case, the number of clock cycles to perform a conversion is 2^{N+1}. Thus for a 16-bit converter with a clock frequency equal to 1 MHz, the worst-case conversion rate is around 7.6 Hz .

Finally, it should be mentioned that by a careful choice for T_1, certain frequency components superimposed on the input signal can be significantly attenuated. Note that this converter effectively "integrates and dumps" the input signal. If we relax the assumption that V_{in} is constant in (17.2), we see that before being converted V_{in} is actually integrated over a time window T_1 in duration,

$$\int_0^{T_1} \frac{V_{in}(\tau)}{R_1 C_1} d\tau \tag{17.9}$$

which is equivalent to convolution by a rectangular time function. Since the Fourier transform of a rectangular pulse is a "sin(x)/x" type response, we have an effective input filter with a transfer function,

$$|H(f)| = \left| \frac{\sin(\pi f T_1)}{(\pi f T_1)} \right| \tag{17.10}$$

Therefore, integrating converters have a low-pass response, with nulls at all integer multiples of $f = 1/T_1$.

Key Point: *Integrating converters inherently filter the input with a sin x/x frequency response.*

EXAMPLE 17.1

If the input signal, V_{in}, is a dc level with power line noise of 60 Hz superimposed on it, choose T_1 to filter out the power line noise.

Solution

Write V_{in} as

$$V_{in} = V_{in(ideal)} + V_{in(60\ Hz)} \tag{17.11}$$

where $V_{in(ideal)}$ is the desired dc signal level and $V_{in(60\ Hz)}$ is the interfering 60 Hz noise; or mathematically,

$$V_{in(60\ Hz)} = A \sin(120\pi t + \phi) \tag{17.12}$$

where A and ϕ are arbitrary magnitude and phase values. Now substituting this relationship for V_{in} into (17.2), we have

$$V_x(T_1) = -\int_0^{T_1} \frac{(-V_{in})}{R_1 C_1}\, d\tau = -\int_0^{T_1} \frac{(-V_{in(ideal)})}{R_1 C_1}\, d\tau - \int_0^{T_1} \frac{(-V_{in(60\ Hz)})}{R_1 C_1}\, d\tau \tag{17.13}$$

However, the last term in (17.13) can be shown to equal zero when T_1 is an integer multiple of $1/(60\ Hz)$ (i.e., 16.67 ms). In this way, the peak value, $V_x(T_1)$, remains correct so that the conversion is performed without error. Note that for this same value of T_1, the harmonics of 60 Hz are also suppressed (i.e., 120 Hz, 180 Hz, 240 Hz, etc.). The converter's frequency response is shown in Fig. 17.3 with $T_1 = 1/(60\ Hz)$. Note that full suppression is achieved at harmonics of 60 Hz, as expected.

In fact, other frequencies are also attenuated but not fully suppressed as are the harmonics of $1/T_1$. Note the –20 dB/decade slope in Fig. 17.3, so that higher frequencies are attenuated more.

EXAMPLE 17.2

It is desired to build a 16-bit two-slope integrating A/D such that a maximum input signal of $V_{in} = 3V$ results in the peak voltage of V_x being 4V. In addition, input noise signals at 50 Hz and harmonics should be significantly attenuated. Find the required RC time constant and clock rate. Also, find the attenuation of a noise signal around 1 kHz superimposed on the input signal.

Solution

Since 50 Hz and harmonics are to be rejected, we choose

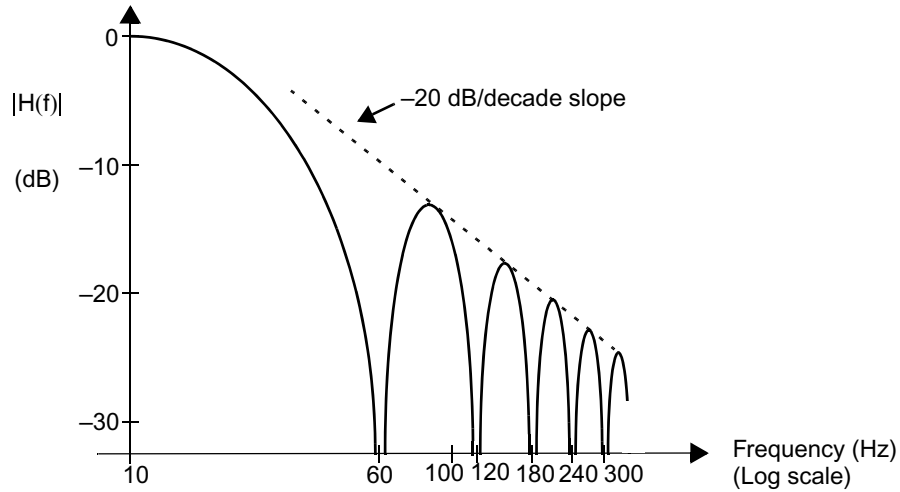


Fig. 17.3 Magnitude response of the effective input filter for an integrating-type converter with $T_1 = 1/(60\ \text{Hz})$

$$T_1 = \frac{1}{50} = 20\ \text{ms} \tag{17.14}$$

Thus, for a 16-bit converter, we require a clock frequency of

$$f_{clk} = \frac{1}{T_{clk}} = \frac{2^{16}}{T_1} = 3.28\ \text{MHz} \tag{17.15}$$

To find the RC time constant needed, we note that at the end of phase (I), V_x is given by

$$V_x = \frac{V_{in} T_1}{R_1 C_1} \tag{17.16}$$

and using the values of $V_x = 4$ V, $V_{in} = 3$ V, and $T_1 = 20$ ms results in

$$R_1 C_1 = 15\ \text{ms} \tag{17.17}$$

Finally, the attenuation of a 1-kHz signal is infinite since it is a harmonic of 50 kHz. However, as seen in (17.10), attenuation is reduced halfway between harmonics, so we find the gain for an input signal at 975 Hz to be

$$|H(f)| = \left| \frac{\sin(\pi \times 975\ \text{Hz} \times 20\ \text{ms})}{\pi \times 975\ \text{Hz} \times 20\ \text{ms}} \right| = 16.3 \times 10^{-3} \tag{17.18}$$

which implies the attenuation is 36 dB.

17.2 SUCCESSIVE-APPROXIMATION CONVERTERS

Successive-approximation A/D converters are one of the most popular approaches for realizing A/D converters due to their amazing versatility. They can provide reasonably quick conversion time, or they can be used for

relatively high accuracy, and can operate with very low power in either case. Fundamentally, these benefits arise because successive-approximation converters require only modest circuit complexity, in the simplest cases requiring only a single comparator, a bank of capacitors with switches, and a small amount of digital control logic.

To understand the basic operation of successive-approximation converters, knowledge of the search algorithm referred to as a "binary search" is helpful. As an example of a binary search, consider the game of guessing a random number from 1 to 128 where one can ask only questions that have a "yes/no" response. The first question might be, "Is the number greater than 64?" If the answer is yes, then the second question asks whether the number is greater than 96. However, if the first answer is no, then the second question asks whether the number is greater than 32. The third question divides the search space in two once again and the process is repeated until the random number is determined. In general, a binary search divides the search space in two each time, and the desired data can be found in N steps for a set of organized data of size 2^N.

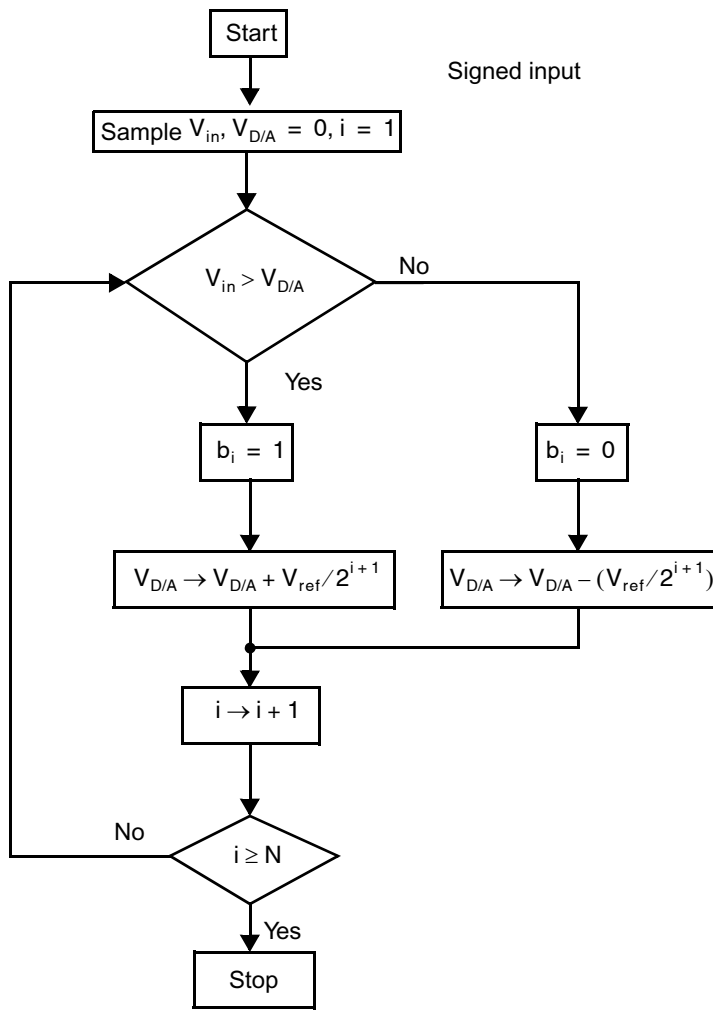


Fig. 17.4 Flow graph for the successive-approximation approach.

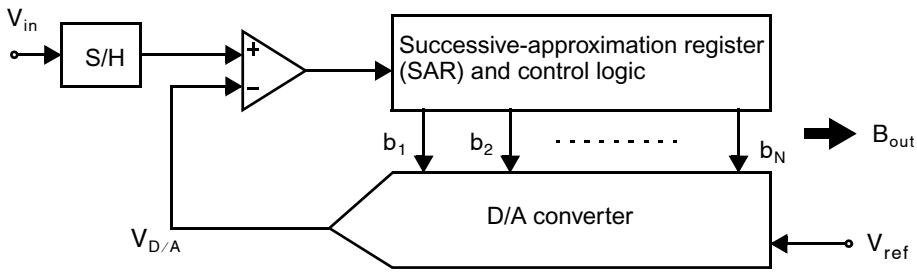


Fig. 17.5 D/A converter-based successive-approximation converter.

Key Point: *Successive-approximation A/D converters are very versatile, capable of moderately high speed or accuracy with relatively low power. They are relatively simple circuits, in the simplest cases requiring only a single comparator, a bank of capacitors and switches, and a small digital logic circuit. Their biggest drawback is that they operate iteratively and therefore require many clock cycles to perform a single conversion.*

Successive-approximation converters apply a binary search algorithm to determine the closest digital word to match an input signal. Specifically, in the first period, after possibly the reset period, the MSB, b_1, is determined. In the second period, the next bit, b_2, is determined, followed by b_3 and so on until all N bits are determined. Thus, in its most straightforward implementation, a successive-approximation converter requires N clock cycles to complete an N-bit conversion. A flow graph for a *signed* conversion using a successive-approximation approach is shown in Fig. 17.4. Here, the signed output is in offset-binary coding and the input signal is assumed to be within $\pm 0.5\ V_{ref}$. The flow graph for a *unipolar* conversion is only slightly different and is left as an exercise for the reader. The major drawback of successive-approximation converters is that, because their principle of operation is an iterative search, they require multiple clock cycles for each input conversion. This limits their conversion frequency to far below the circuit's maximum clock frequency, particularly when high resolution is sought since more iterations are required.

17.2.1 D/A-Based Successive Approximation

The block diagram for a unipolar successive-approximation A/D converter that uses a D/A converter is shown in Fig. 17.5. The successive-approximation register (SAR) and control logic are entirely digital and perform the necessary binary search. At the end of the conversion, the digital value in the SAR results in the voltage $V_{D/A}$ being within $0.5\ V_{LSB}$ of the input signal. With this type of architecture, the D/A converter typically determines the accuracy and speed of the A/D converter. Note that a sample and hold is required at the input so that the value to be converted does not change during the conversion time.

EXAMPLE 17.3

Consider the case where $V_{ref} = 1$ V, $V_{in} = 0.354$ V, and a 3-bit conversion is performed. Find intermediate D/A values and the final output.

Solution

In this case, $V_{LSB} = 0.125$ V.

In cycle 1: $B_{out} = 100$ so that $V_{D/A} = 0.5$ V. Since $V_{in} < V_{D/A}$, b_1 is set to 0.
In cycle 2: $B_{out} = 010$ so that $V_{D/A} = 0.25$ V. Since $V_{in} > V_{D/A}$, b_2 is set to 1.
In cycle 3: $B_{out} = 011$ so that $V_{D/A} = 0.375$ V. Since $V_{in} < V_{D/A}$, b_3 is set to 0.

Therefore, the resulting output is the last value of B_{out}, which is 010. Although the final quantization error is 0.831 V_{LSB}, it is greater than $\pm 0.5\ V_{LSB}$ because we have not accounted for the 0.5 LSB offset as discussed at the beginning of this chapter.

17.2.2 Charge-Redistribution A/D

The straightforward approach of using a separate D/A converter and setting it equal to the input voltage (within one LSB) can be modified to the flow graph shown in Fig. 17.6. Here, the error signal V equals the difference between the input signal, V_{in}, and the D/A output, $V_{D/A}$. As a result, V is always compared to ground, as seen at the top of the flow graph, and the goal is to set this error difference within one LSB of zero.

One of the first switched-capacitor analog systems using this approach is a charge-redistribution MOS A/D converter [McCreary, 1975]. With such a converter, the sample and hold, D/A converter, and the difference portion of the comparator are all combined into a single circuit. The unipolar case is shown in Fig. 17.7 and operates as follows:

1. *Sample mode:* In the first step, all the capacitors are charged to V_{in} while the comparator is being reset to its threshold voltage through s_2, In this step, note that the capacitor array is performing the sample-and-hold operation.
2. *Hold mode:* Next, the comparator is taken out of reset by opening s_2, and then all the capacitors are switched to ground. This causes V_x, which was originally zero, to change to $-V_{in}$, thereby holding the input signal, V_{in}, on the capacitor array. (This step is sometimes merged with the first bit time during bit cycling.) Finally, s_1 is switched so that V_{ref} can be applied to the capacitor array during bit cycling.
3. *Bit cycling:* Next, the largest capacitor (i.e., the $16C$ capacitor in this example) is switched to V_{ref}. V_x now goes to $(-V_{in} + V_{ref}/2)$. If V_x is negative, then V_{in} is greater than $V_{ref}/2$, and the MSB capacitor is left connected to V_{ref}. Also b_1 is considered to be a 1. Otherwise, the MSB capacitor is reconnected to ground and b_1 is taken to be 0. This process is repeated N times, with a smaller capacitor being switched each time, until the conversion is finished.

To get an exact division by two, note that an additional unit capacitor of size C has been added so that the total capacitance is $2^N C$ rather than $(2^N - 1)C$. Also, the capacitor *bottom plates* should be connected to the V_{ref} side, not to the comparator side, to minimize the parasitic capacitance at node V_x. Although parasitic capacitance at V_x does not cause any conversion errors with an ideal comparator, it does attenuate the voltage V_x.

A signed A/D conversion can be realized by adding a $-V_{ref}$ input. If V_x is less than zero at the first step, then proceed as in the unipolar case using V_{ref}. Otherwise, if V_x is greater than zero, use $-V_{ref}$ and test for V_x greater than zero when deciding whether to leave the capacitors connected to $-V_{ref}$ or not at each bit cycling.

Key Point: *The accuracy of successive-approximation converters is often determined by that of the D/A converter. In the case of charge-redistribution converters, this role is played by a bank of comparators whose matching is important.*

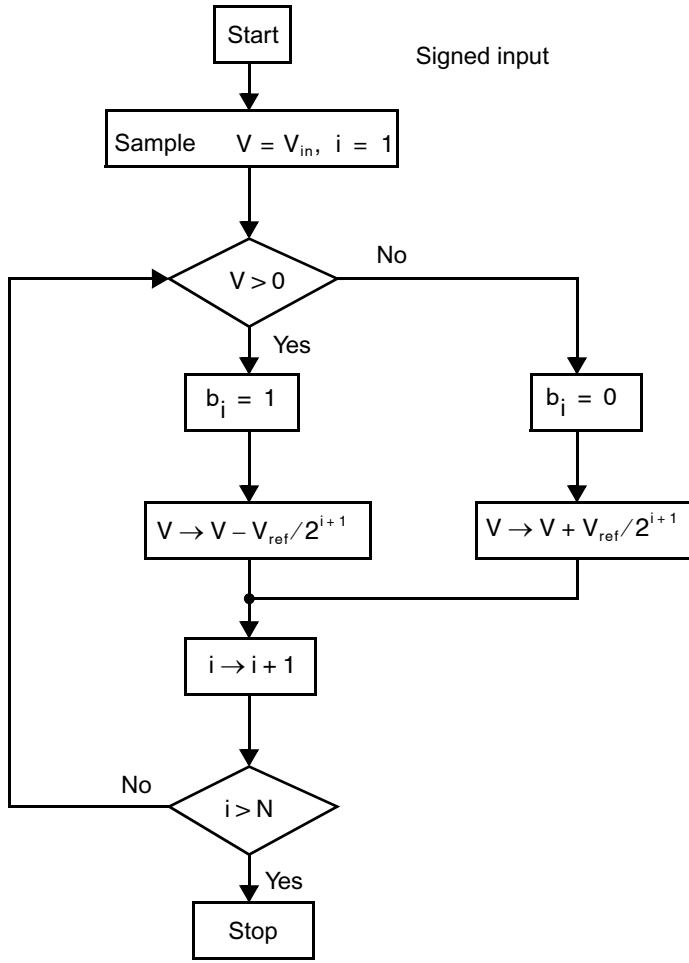


Fig. 17.6 Flow graph for a modified successive approximation (divided remainder).

EXAMPLE 17.4

Find intermediate node voltages at V_x during the operation of the 5-bit charge-redistribution converter shown in Fig. 17.7 when $V_{in} = 1.23$ V and $V_{ref} = 5$ V. Assume a parasitic capacitance of $8C$ exists on the node at V_x.

Solution

First, $V_x = 0$ during sample mode. Next, during hold mode, all capacitors are switched and the charge on V_x is shared between the $32C$ total converter capacitance and the parasitic $8C$ capacitance, resulting in

$$V_x = \frac{32}{32 + 8} \times -V_{in} = -0.984 \text{ V} \tag{17.19}$$

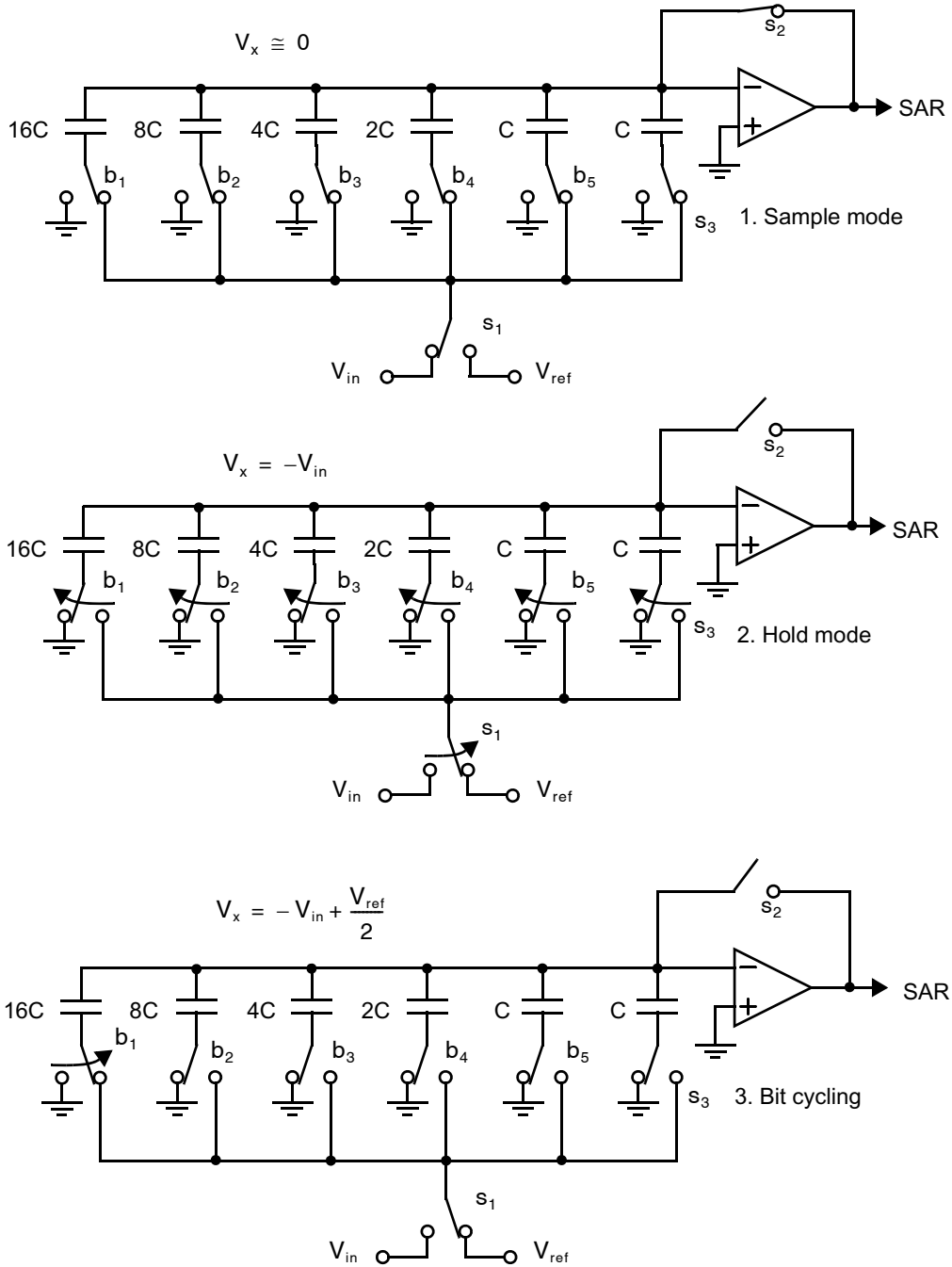


Fig. 17.7 A 5-bit unipolar charge-redistribution A/D converter.

During the first bit cycling, b_1 is switched, resulting in

$$V_x = -0.984 + \frac{16}{40} \times 5 \text{ V} = 1.016 \text{ V} \tag{17.20}$$

Since this result is greater than zero, switch b_1 is reversed back to ground and V_x returns to –0.984 V. Thus, $b_1 = 0$.

Next, b_2 is switched, and we have

$$V_x = -0.984 + \frac{8}{40} \times 5 \text{ V} = 0.016 \text{ V} \tag{17.21}$$

It is also greater than zero, so $b_2 = 0$, and V_x is set back to –0.984 V by switching b_2 back to ground.

When b_3 is next switched, we have

$$V_x = -0.984 + \frac{4}{40} \times 5 = -0.484 \text{ V} \tag{17.22}$$

Since this result is less than zero, $b_3 = 1$, and switch b_3 is left connected to V_{ref}.

Next, b_4 is switched, resulting in

$$V_x = -0.484 + \frac{2}{40} \times 5 = -0.234 \text{ V} \tag{17.23}$$

which is also less than zero, so $b_4 = 1$, and this switch is left connected to V_{ref}.

Finally, b_5 is switched, resulting in

$$V_x = -0.234 + \frac{1}{40} \times 5 = -0.109 \text{ V} \tag{17.24}$$

which is also less than zero, so $b_5 = 1$.

Therefore, the output is given by $B_{out} = 00111$ and voltage V_x is within a V_{LSB} of ground ($V_{LSB} = 5/32$ V).

The same structure as the unipolar case can be used to realize a signed A/D conversion while using only a single V_{ref} if a slightly modified switching arrangement is used. Referring to Fig. 17.8, and assuming V_{in} is between $\pm V_{ref}/2$, the conversion proceeds as follows:

1. *Sample mode:* In the first step, all the capacitors, except for the largest capacitor, are charged to V_{in} while the comparator is being reset to its threshold voltage. For the signed case, the largest capacitor is now connected to $V_{ref}/2$.
2. *Hold mode:* Next, the comparator is first taken out of reset, and then all the capacitors, except the largest one, are switched to ground. This causes V_x, which was originally zero, to change to $-V_{in}/2$. At the end of this step, the sign of the input signal is determined by looking at the comparator output.
3. *Bit cycling:* Next, the largest capacitor (i.e., the $16C$ capacitor in this example) is switched to ground *if, and only if,* V_x is larger than zero (i.e., when V_{in} is less than zero). Specifically, if V_x is less than zero, then V_{in} is positive and b_1 is set to 1, and conversion proceeds as in the unipolar case, starting with b_2, until conversion is completed. However, if V_x is larger than zero, b_1 is set to 0, the largest capacitor is switched to ground, causing V_x to become $-V_{in}/2 - V_{ref}/4$ (which is a negative value), and conversion proceeds as in the unipolar case, starting with b_2. Once conversion is completed, some digital recoding may be required to obtain the desired output code.

This approach for realizing signed A/Ds has the disadvantage that V_{in} has been attenuated by a factor of two, which makes noise more of a problem for high-resolution A/Ds. Also, any error in the MSB capacitor now causes both an offset and a sign-dependent gain error. The latter causes integral nonlinearity errors.

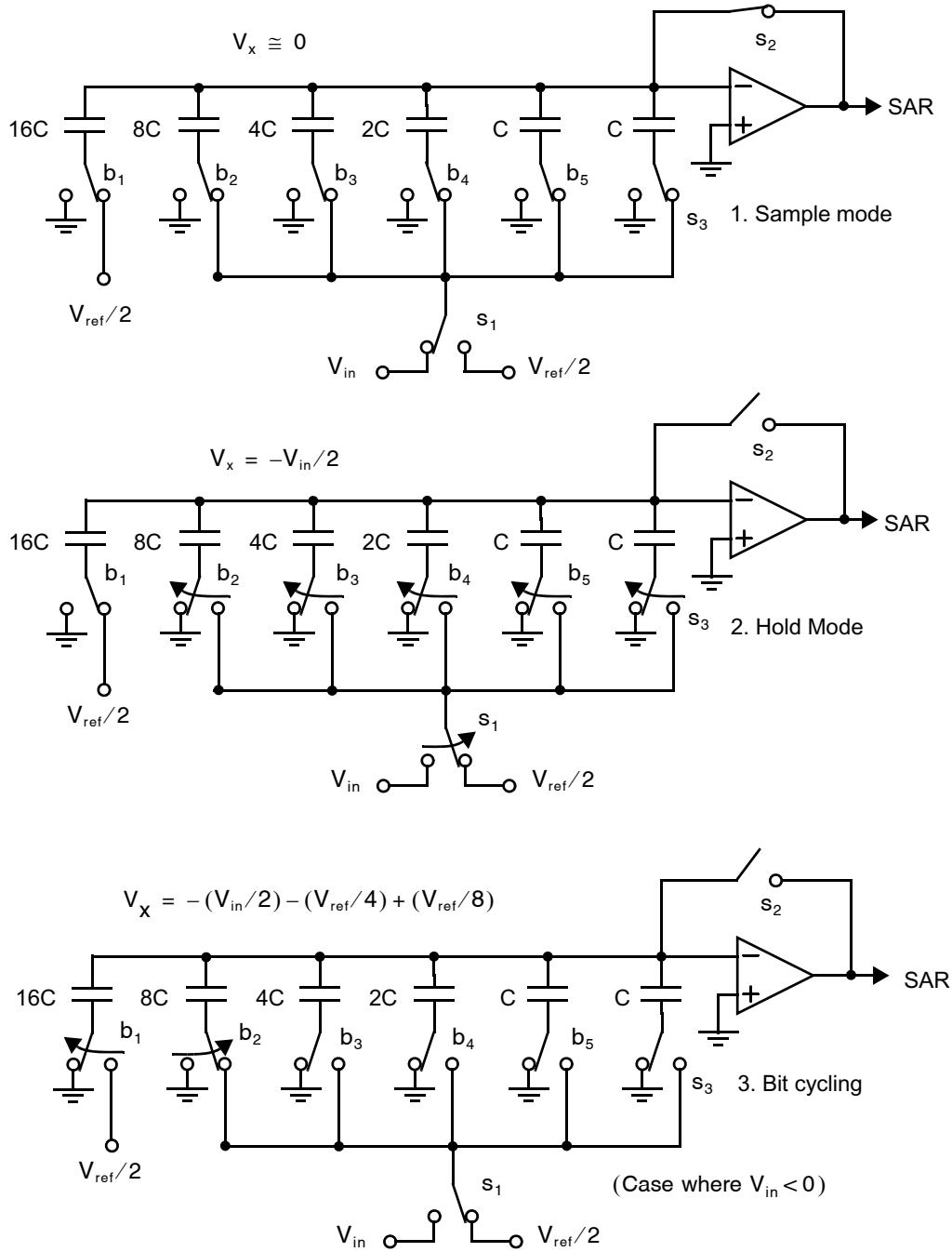


Fig. 17.8 A 5-bit signed charge-redistribution A/D converter.

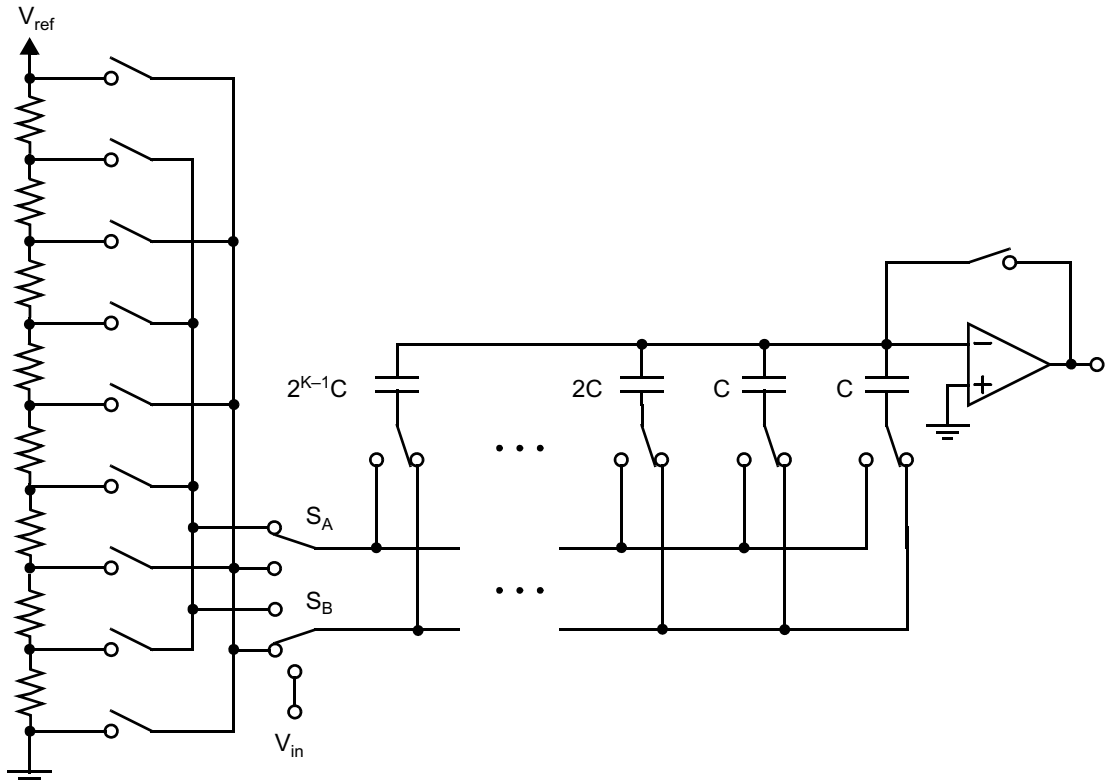


Fig. 17.9 Resistor–capacitor hybrid A/D converter.

17.2.3 Resistor–Capacitor Hybrid

A combination of the resistor-string and capacitor-array approaches in a hybrid A/D converter, similar to the hybrid D/A converter [Fotouhi, 1979], is shown in Fig. 17.9.

The first step is to charge all the capacitors to V_{in} while the comparator is being reset. Next, a successive-approximation conversion is performed to find the two adjacent resistor nodes that have voltages larger and smaller than V_{in}. One bus will be connected to one node while the other is connected to the other node. All of the capacitors are connected to the bus having the lower voltage. A successive approximation using the capacitor-array network is then done. Starting with the largest capacitor, a capacitor is switched to the adjacent resistor-string node having a larger voltage. If the comparator output is a 1, it is switched back and b_i is a 0. Otherwise, the switch is left as is and b_i is a 1. Since the resistor string is monotonic, this type of converter is guaranteed monotonic if the capacitor array is monotonic.

17.2.4 Speed Estimate for Charge-Redistribution Converters

The major limitation on speed with charge redistribution converters is often due to the RC time constants of the capacitor array and switches. To estimate this time, consider the simplified model of a capacitor array being reset,

as shown in Fig. 17.10. Here, R, R_{s1}, and R_{s2} represent the switch-on resistances of the bit line, $s1$ and $s2$ switches, respectively. Although this circuit is easily simulated using SPICE to find its settling time, it is useful to have a rough estimate of the charging time to speed up the design process. As in Section 16.1, the zero-value time-constant approach [Sedra, 1991] can be used to obtain an estimate of the high-frequency time constant by summing individual time constants due to each capacitor. For example, the individual time constant due to the capacitance $2C$ equals $(R_{s1} + R + R_{s2})2C$. Following such an approach, the zero-value time constant for the circuit shown in Fig. 17.10 is equal to

$$\tau_{eq} = (R_{s1} + R + R_{s2})2^N C \tag{17.25}$$

For better than 0.5-LSB accuracy, we need

$$e^{-T/\tau_{eq}} < \frac{1}{2^{N+1}} \tag{17.26}$$

where T is the charging time. This equation can be simplified to

$$T > \tau_{eq}(N+1)\ln(2) = 0.69(N+1)\tau_{eq} \tag{17.27}$$

It has been observed that (17.27) gives results about 30 percent higher than those obtained by simulating the actual RC network shown in Fig. 17.10. So while (17.27) can be used to roughly determine the maximum sampling rate, the final design should be simulated using accurate transistor models. Finally, although this result is not the same for all charge-redistribution A/D converters (for example, the size of the switches going to the larger capacitors may be increased to reduce their on resistance), the basic approach can be modified for most charge-redistribution A/D converters.

17.2.5 Error Correction in Successive-Approximation Converters

With the best matching accuracy of on-chip elements being about 0.1 percent, one is limited to successive-approximation converters having 10-bit accuracy specifications without some sort of calibration. One error-correction technique that has been used to obtain 16-bit linear converters is shown in Fig. 17.11 [Lee, 1984]. In this approach, the MSB array is realized using binary-weighted capacitors that determine, for example, the first 10 bits.

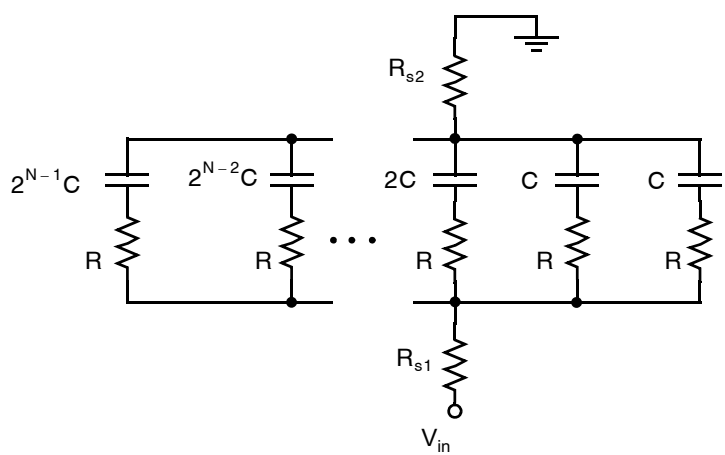


Fig. 17.10 Simplified model of a capacitor array during the sampling time.

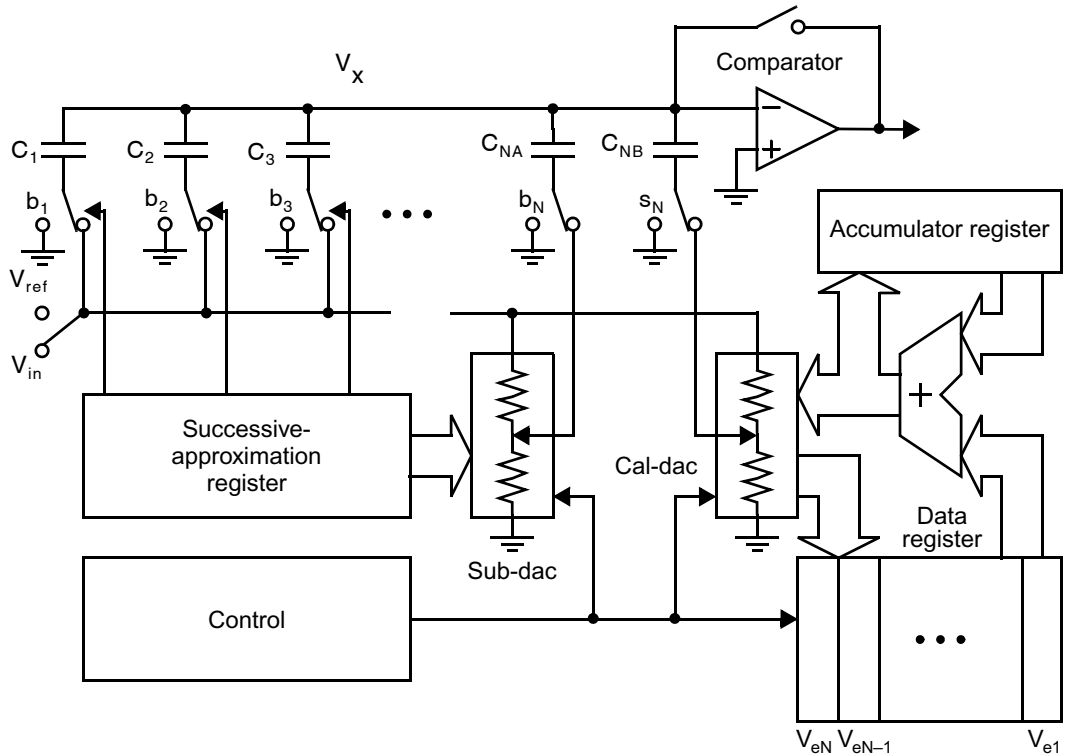


Fig. 17.11 A charge-redistribution A/D converter with error correction.

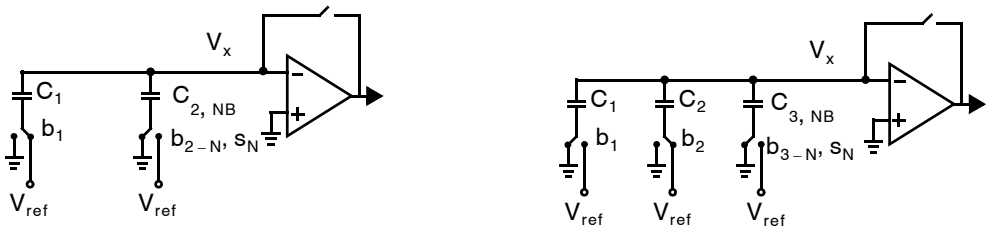
For a 16-bit converter, the final 6 bits are determined using an additional capacitor and a resistor string referred to as a *sub-dac.* Although this combination of an MSB capacitor array and an LSB resistor string is not inherently monotonic, it can be easily autocalibrated at start-up by adding a second resistor string referred to as a *cal-dac.*

Calibration is done by measuring the errors of each capacitor, starting with the largest capacitor, calculating the correction terms required, and then storing the correction terms in a data register as DV_{ei}. During a regular successive approximation operation, whenever a particular capacitor is used, its error is cancelled by adding the value stored in the data register to that stored in an accumulator register, which contains the sum of the correction terms for all of the other capacitors currently connected to V_{ref}. No correction terms are measured for the resistor sub-dac; its accuracy is not critical since it only determines the remaining LSBs.

The error terms are found starting with the MSB capacitor, C_1. It is connected to ground and all of the other capacitors are connected to V_{ref} while the comparator is reset. Next, the comparator is taken out of reset mode, all of the other capacitors are switched to ground, and C_1 is switched to V_{ref}. Defining C_{total} to be the sum of all the capacitors in the array, the ideal value of C_1 is $C_{total}/2$. However, in practice C_1 varies from its ideal value, and thus, C_1 can be written as

$$C_1 \equiv (C_{total}/2) + \Delta C_1 \tag{17.28}$$

where ΔC_1 is the capacitance error (either positive or negative). As a result, the simplified model shown in Fig. 17.12(*a*) occurs after the switches are reversed. Here, the remaining capacitance equals $(C_{total}/2) - \Delta C_1$ to make the total capacitance correct. With the above switch operation, V_x would remain zero if ΔC_1 equals zero. However, when ΔC_1 is not zero, the resulting voltage, V_x, is twice the error voltage, defined as V_{e1}, that would be introduced during a normal successive approximation when C_1 alone is switched to V_{ref}. A digital representation



$$C_1 = \left(\frac{C_{total}}{2}\right) + \Delta C_1 \qquad C_2 = \left(\frac{C_{total}}{4}\right) + \Delta C_2$$

$$C_{2,NB} = \left(\frac{C_{total}}{2}\right) - \Delta C_1 \qquad C_{3,NB} = \left(\frac{C_{total}}{4}\right) - \Delta C_1 - \Delta C_2$$

(*a*) (*b*)

Fig. 17.12 Equivalent models for determining capacitance errors. (*a*) During calibration for C_1. (*b*) During calibration for C_2.

of V_{e1}, defined as DV_{e1}, is obtained by doing a successive approximation using the cal-dac shown in Fig. 17.11 and then dividing the resulting digital value by 2 to obtain DV_{e1}. This digital correction term, DV_{e1}, is stored in the data register for use during regular conversion.

To obtain the similar correction term for C_2, defined as DV_{e2}, the model in Fig. 17.12(*b*) is used. Here, the switch procedure is to always leave the b_1 switch grounded while the comparator is reset, the b_2 switch is grounded, and the remaining switches b_{3-N}, s_N are connected to V_{ref}. Next, the b_2 and b_{3-N}, s_N switches are all reversed (i.e., b_2 set to V_{ref} and the others grounded). Note, however, that even if ΔC_2 equals zero, the measured value of V_x will be equivalent to $-DV_{e1}$, since $-\Delta C_1$ is assumed to be part of $C_{3,NB}$. Therefore, the error voltage that would be due only to C_2 can be calculated by digitizing V_x, as before, but then subtracting off DV_{e1} using digital circuitry. Mathematically,

$$DV_{e2} = \frac{1}{2}(DV_{x2} - DV_{e1}) \tag{17.29}$$

A similar procedure is used for the other capacitors. Specifically, DV_{xi} is found by connecting all capacitors smaller than C_i to V_{ref}, and connecting C_i and all larger capacitors to ground while the comparator is reset. Next, the comparator is taken out of reset mode, all capacitors smaller than C_i are switched to ground, C_i is switched to V_{ref}, and all capacitors larger than C_i are left connected to ground. DV_{xi} is then found using the cal-dac and successive approximation. DV_{ei} is finally calculated using the formula

$$DV_{ei} = \frac{1}{2}\left(DV_{xi} - \sum_{j=1}^{i-1} DV_{ej}\right) \tag{17.30}$$

and stored in the ith word of the data register.

During a regular conversion, normal successive approximation is performed with the MSB capacitor array; however, appropriate correction voltages are either added or subtracted using the cal-dac and C_{NB} capacitor. The proper correction voltage is determined through the use of a digital accumulator that stores the sum of all digital errors of those MSB capacitors deemed to be a 1 (i.e., connected to V_{ref}). In other words, when the ith bit is being tested, DV_{ei} is added to the digital accumulator that is driving the cal-dac. If the ith bit is determined to be a 0, then the digital accumulator returns to its previous value; otherwise, it maintains its new accumulated value.

Finally, normal successive approximation is performed using the sub-dac to determine the final LSBs. With this approach, a digital addition must be performed during each bit cycle, and a small amount of digital RAM is required (10 bytes in the case of a 10-bit MSB array).

Finally, it should be mentioned that similar error correction techniques have also been described to account for capacitor inaccuracies (for example, [Tan, 1990]).

17.2.6 Multi-Bit Successive-Approximation

The successive-approximation converters we have seen so far divide their search space into two on each clock cycle by performing a single comparison. The search can be accelerated by performing multiple comparisons on each clock cycle, and dividing the search space up into smaller regions. For example, if we are trying to guess the value of a random number in the range 1 to 128, in addition to asking the question "Is the number greater than 64?," imagine that we can simultaneously ask "Is the number greater than 32?" and "Is the number greater than 96?," With the answers to all three of these questions, we can divide the search space into a quarter: We will know if the number is in the range 1–32, 33–64, 65–96, or 97–128. Hence, we have effectively performed the first two iterations of a binary search. In an analog circuit, this can be done by operating three comparators in parallel. The result is that we require only one-half the number of clock cycles to perform an A/D conversion, hence potentially doubling the converter's sampling rate. However, note that doing so required three comparators (and likely three capacitor banks) instead of only one. A flow graph depicting multi-bit successive-approximation is shown in Fig. 17.13. Naturally, the approach can be extended to perform any number of bits per iteration, but in order to find L bits on each clock cycle, the number of comparisons required $2^L - 1$ grows exponentially. Therefore, the number of bits is usually restricted to at most a few.

17.3 ALGORITHMIC (OR CYCLIC) A/D CONVERTER

Key Point: *Algorithmic converters are similar to successive-approximation converters, except that instead of an accurate D/A converter or capacitor bank, these converters require an accurate gain of two in order to perform the iterative search for a digital code that represents the analog input.*

An algorithmic converter operates in much the same way as a successive-approximation converter. However, whereas a successive-approximation converter halves the reference voltage in each cycle, an algorithmic converter doubles the error voltage while leaving the reference voltage unchanged. The flow graph for a signed algorithmic conversion is shown in Fig. 17.14.

17.3.1 Ratio-Independent Algorithmic Converter

The block diagram for an algorithmic converter is shown in Fig. 17.15 [McCharles, 1977; Li, 1984]. This converter requires a small amount of analog circuitry because it repeatedly uses the same circuitry to perform its conversion cyclically in time.

One of the difficulties in realizing a high-precision algorithmic converter is building an accurate multiply-by-two gain amp. Fortunately, it is possible to realize the gain amp so that it does not rely on any capacitor matching if four clock cycles are taken for the multiply-by-two operation. The operation of the multiply-by-two gain amp is shown in Fig. 17.16. Although this gain-amp circuitry is shown using single-ended circuits for simplicity, fully differential circuits are normally used.

The basic idea of this gain amp is to sample the input signal twice using the same capacitor. During the second sampling, the charge from the first capacitor is stored on a second capacitor whose size is unimportant. After the second sampling, both charges are recombined into the first capacitor which is then connected between the opamp input and output.

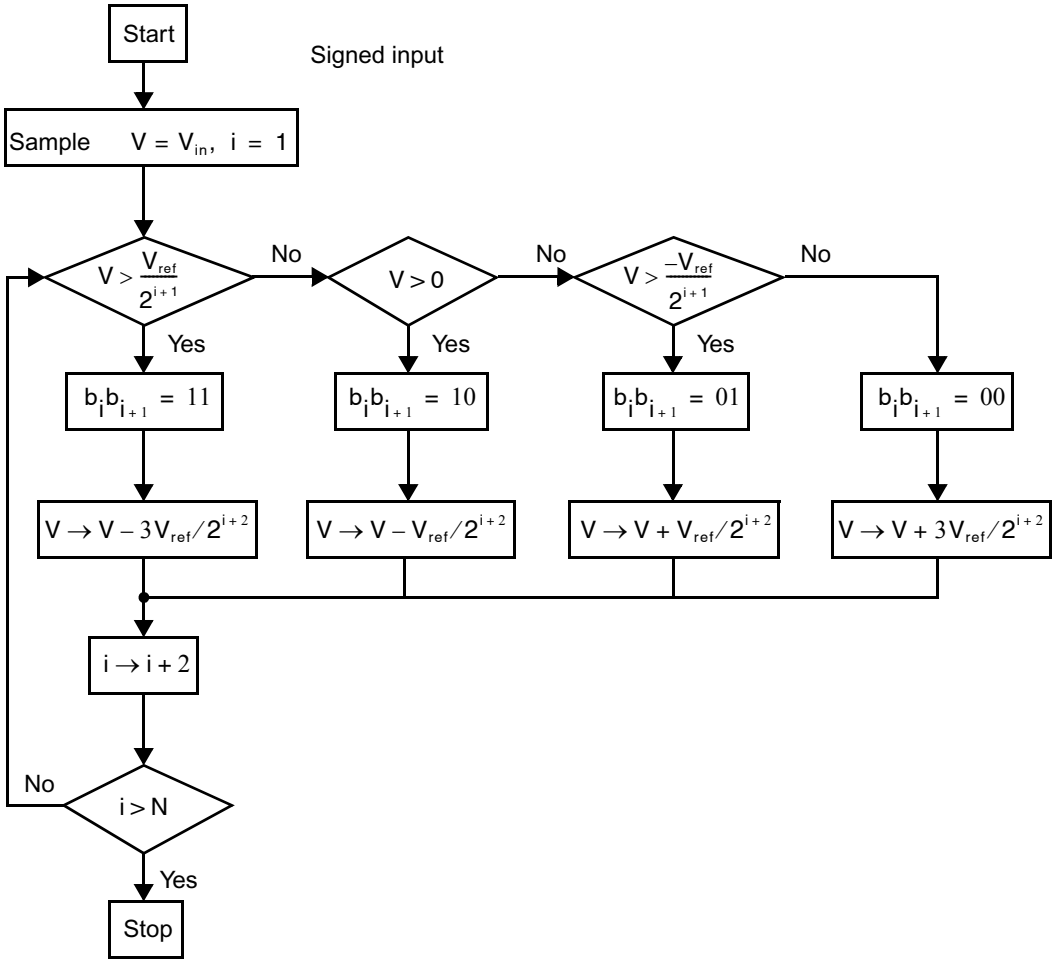


Fig. 17.13 Flow graph for a 2-bit per iteration successive approximation.

EXAMPLE 17.5

Consider the multiply-by-two gain circuitry shown in Fig. 17.16. Assuming the opamp has an input offset designated as V_{off}, find the values of V_{C1}, V_{C2}, and V_{out} at the end of each of the phases shown.

Solution

During phase one, the opamp output, V_{out}, is connected to the negative opamp input, resulting in

$$V_{out} = V_{off} \tag{17.31}$$

and the voltages across the two capacitors are

$$V_{C1} = V_{err} - V_{off} \tag{17.32}$$

$$V_{C2} = 0 - V_{off} = -V_{off} \tag{17.33}$$

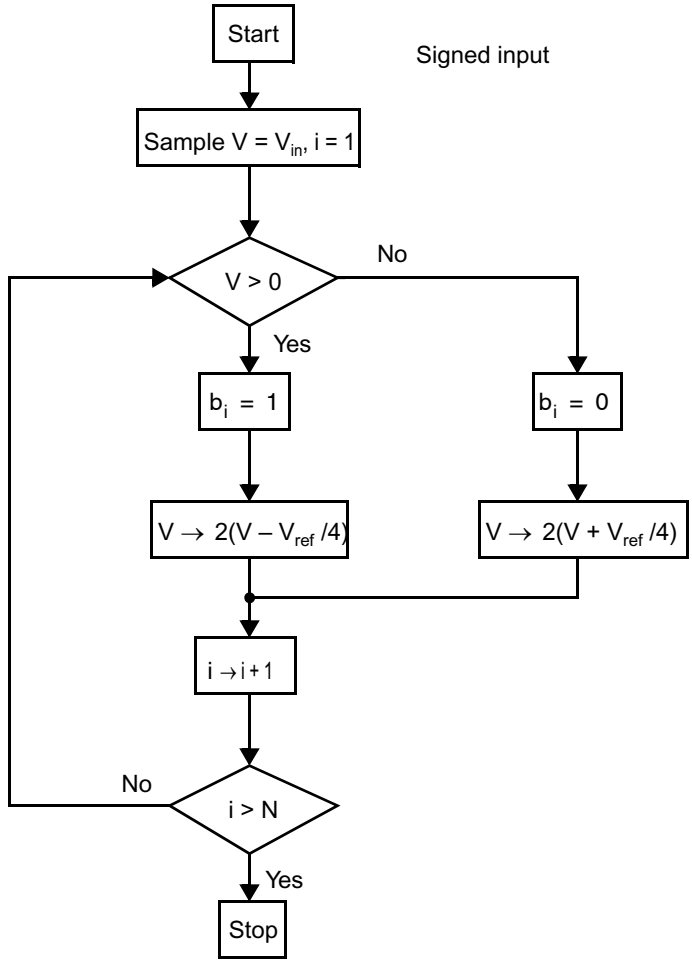


Fig. 17.14 Flow graph for the algorithmic and pipelined converter approach.

At the end of phase two, we have

$$V_{C1} = 0 - V_{off} = -V_{off} \tag{17.34}$$

implying that its charge change was

$$\Delta Q_{C1} = C_1(V_{err} - V_{off} - (-V_{off})) = C_1 V_{err} \tag{17.35}$$

All this charge is placed on C_2, resulting in

$$V_{C2} = -V_{off} + \left(\frac{C_1}{C_2}\right) V_{err} \tag{17.36}$$

and $V_{out} = (C_1/C_2)V_{err}$.

At the end of phase three, V_{C2} remains unchanged since one side of it has been opened. Also, $V_{C1} = V_{err} - V_{off}$ and $V_{out} = V_{off}$ as in phase one.

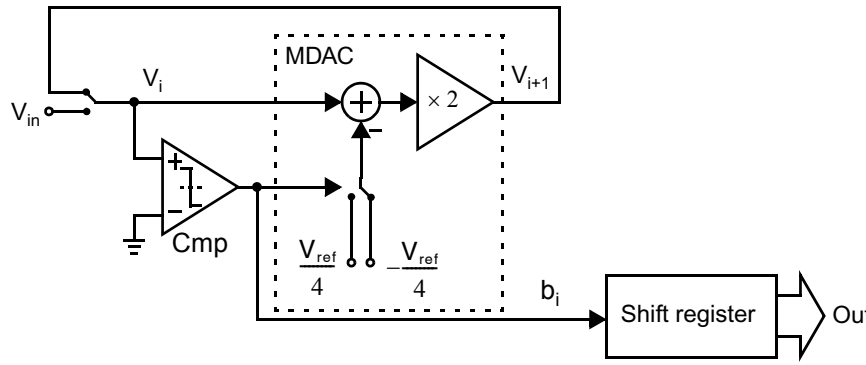


Fig. 17.15 Algorithmic converter.

Finally, at the end of phase four, C_2 is discharged to the same value it had in phase 1, resulting in $V_{C2} = -V_{off}$ and its change in charge being

$$\Delta Q_{C2} = C_2\left(\frac{C_1}{C_2}\right)V_{err} = C_1 V_{err} \tag{17.37}$$

All of this charge is placed back on C_1, resulting in

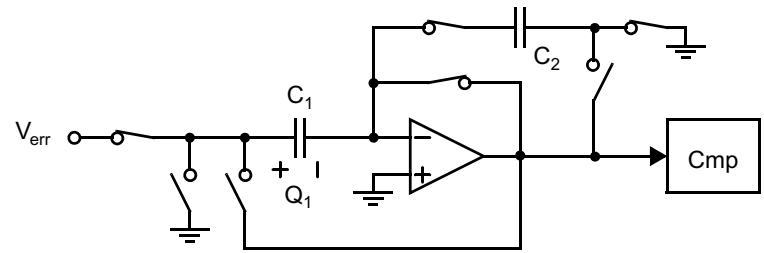
$$V_{C1} = 2\,V_{err} - V_{off} \tag{17.38}$$

and the output voltage being the desired result of $V_{out} = 2\,V_{err}$. Note that this final result is independent of the sizes of C_1, C_2, and the offset value, V_{off}.

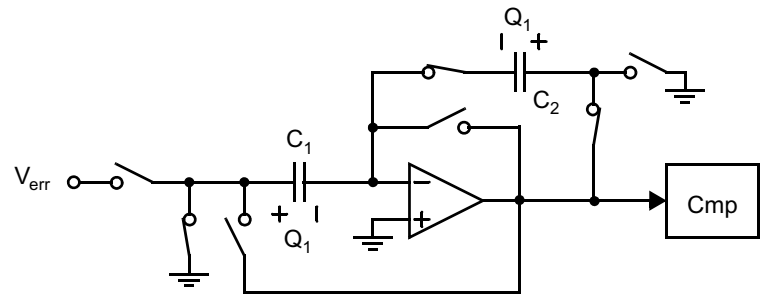
17.4 PIPELINED A/D CONVERTERS

Pipelined converters, like successive-approximation and algorithmic converters, perform an iterative search for a digital code that accurately reflects the analog input signal. However, rather than perform the iterations with a single analog circuit, pipelined converters have a separate analog stage dedicated to performing each iteration. Signal flow in a pipelined converter is diagrammed in Fig. 17.17. All of the analog stages operate on every clock cycle, each operating on a different input sample. Since N iterations are performed simultaneously, the pipelined converter operates N-times faster than algorithmic converters and is capable of outputting one conversion on every clock cycle. Therefore, pipelined converters are generally applied when higher speed is desired than is achievable with algorithmic converters. Although N input samples are being processed in parallel, N clock cycles are required for each input sample to proceed through the entire pipeline. Hence, there is a *latency* of N clock cycles through pipelined converters, just as in algorithmic and successive approximation converters.

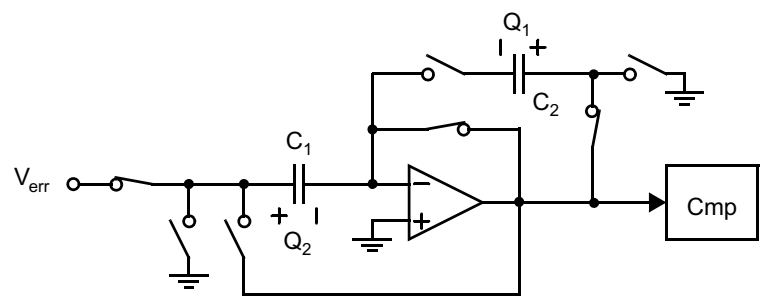
Key Point: *Pipelined A/D converters are similar to algorithmic converters, but employ multiple circuits working on successive input samples simultaneously. Hence, they can complete a conversion on every clock cycle, providing higher throughput than algorithmic converters, but with the same latency.*



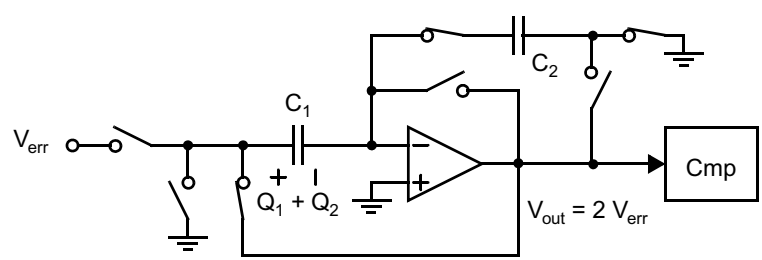
1. Sample remainder and cancel input-offset voltage.



2. Transfer charge Q_1 from C_1 to C_2.



3. Sample input signal with C_1 again, after storing charge Q_1 on C_2.



4. Combine Q_1 and Q_2 on C_1, and connect C_1 to output.

Fig. 17.16 Multiply-by-two gain circuitry for an algorithmic converter that does not depend on capacitor matching.

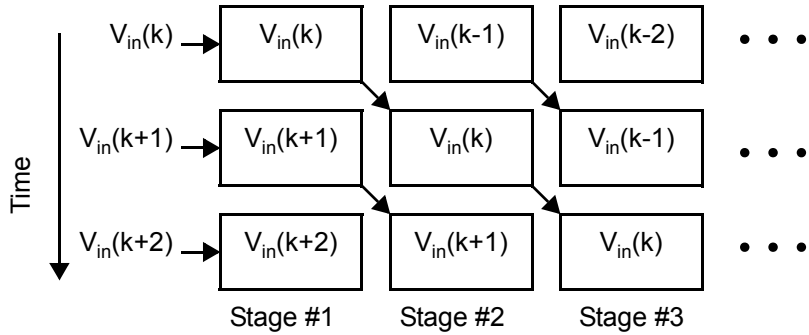


Fig. 17.17 The signal flow in a pipelined A/D converter.

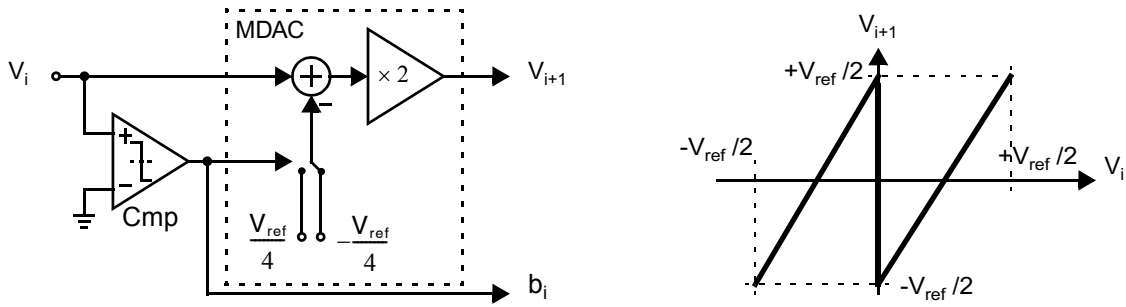


Fig. 17.18 One-bit pipelined converter stage that accepts inputs in the range $-V_{ref}/2 < V_i < V_{ref}/2$.

17.4.1 One-Bit-Per-Stage Pipelined Converter

The operations performed in each stage of a one-bit-per-stage pipelined converter are the same as those for the algorithmic converter shown in Fig. 17.14. Each stage accepts an input from the previous stage, compares it to a reference threshold, and uses a multiplying D/A converter (MDAC) to form a *residue* signal as follows:

$$V_{i+1} = 2[V_i + (b_i - 0.5)V_{ref}/2] \tag{17.39}$$

The block diagram of a single stage is in Fig. 17.18. It's functionality is very similar to that of the algorithmic converter in Fig. 17.15 except rather than returning the resulting Since a single comparison is performed in each stage, each finds a single bit of the result, b_i. However, since each stage is operating on a different input sample, these digital bits must be realigned using shift registers of varying lengths, as shown in Fig. 17.19. The result of the conversion may be interpreted as a N-bit quantized estimate of the original input sample within the range $\pm V_{ref}/2$,

$$\hat{V}_{in} = V_{ref}\sum_{i=1}^{N}(b_i - 0.5)2^{-i} \tag{17.40}$$

Finally, note that the final residue signal V_{N+1} is not used by any subsequent stage so there is no need to generate it. Hence the final MDAC may be omitted and the last stage is simply a 1-bit A/D converter (comparator) generating b_N.

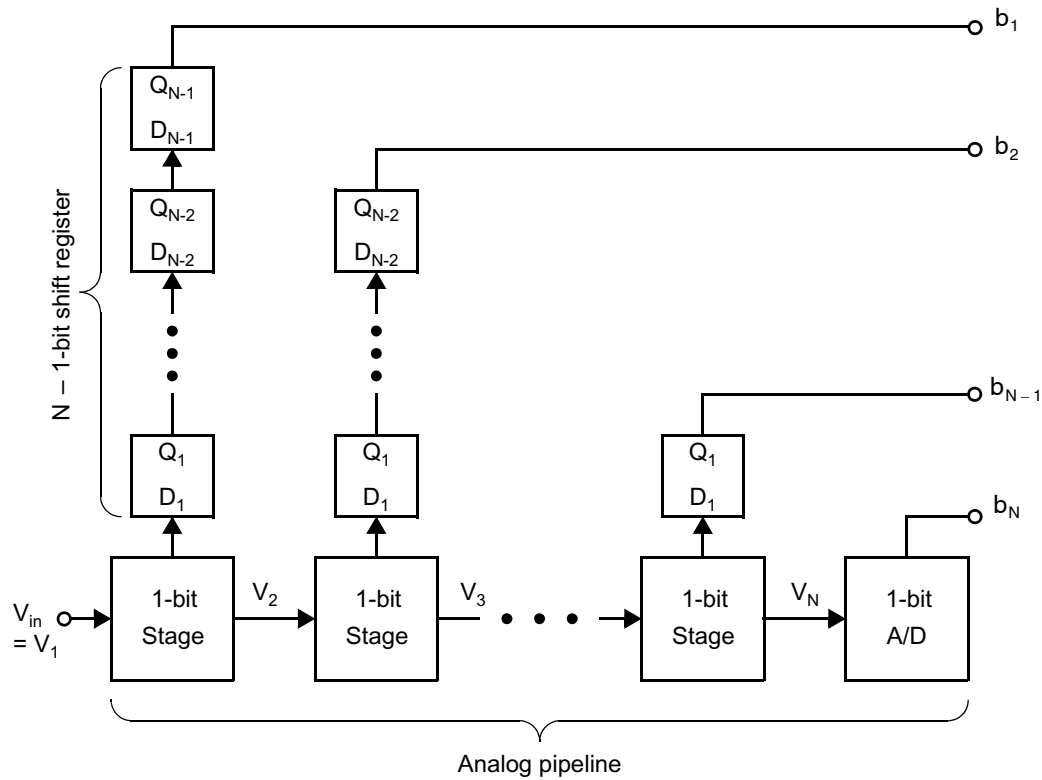


Fig. 17.19 A one-bit-per-stage pipelined A/D converter [Martin, 1981].

EXAMPLE 17.6

Consider the operation of a 3-bit, one-bit-per-stage pipelined converter with $V_{ref} = 1$ V. Find all of the residue voltages, V_i, and the final output digital code for a sampled input voltage of $V_{in} = 240$ mV.

Solution

The input signal is given, $V_1 = V_{in} = 240$ mV. The comparator in the first pipelined stage will output a positive result, $b_1 = 1$. It will, therefore, subtract $V_{ref}/4 = 250$ mV from its input and multiply the result by two.

$$V_2 = 2(V_1 - V_{ref}/4) = -20 \text{ mV}$$

Since $V_2 < 0$, in the next clock cycle the comparator of the second pipelined stage will produce a logic low result, $b_2 = 0$, causing $V_{ref}/4$ to be added for the next residue.

$$V_3 = 2(V_2 + V_{ref}/4) = 460 \text{ mV}$$

Finally, the third stage will output a high logic level resulting in a final output code of $b_1b_2b_3 = 101$ equivalent to a quantized input of $\hat{V}_{in} = 187.5$ mV.

A problem with the one-bit-per-stage pipelined architecture is that any comparator offset results in irreparable errors. In order for a one bit per stage pipelined converter to have 10-mV resolution, the first stage comparator must have better than 10-mV accuracy, the second stage accuracy must be better than 20 mV, the third stage better than 40 mV, and so on. The accuracy requirements diminish as one proceeds along the pipeline since later stages are acting on amplified residue signals, so their impact on the converter's overall accuracy is divided by all of the preceding gain stages when input-referred. Nevertheless, the limitations imposed by comparator offsets mean that the one bit per stage pipeline architecture is rarely used.

EXAMPLE 17.7

Repeat Example 17.6, but this time assume that the comparator in the second stage has an offset of 30 mV.

Solution

As in Example 17.6, the comparator in the first pipelined stage will output a positive result, $b_1 = 1$, resulting in $V_2 = -20$ mV . This time, however, the second stage comparator offset will cause the second pipelined stage to produce a logic high result, $b_2 = 1$, causing $V_{ref}/4$ to be subtracted for the next residue.

$$V_3 = 2(V_2 - V_{ref}/4) = -540 \text{ V}$$

This residue is outside the range $\pm V_{ref}/2$ and therefore results in an excess error in the converter's final result. The final stage will output a logic low resulting in the output code: $b_1b_2b_3 = 110$ or $V_{in} = 375$ mV, a less accurate result than the output produced without offset in Example 17.6.

17.4.2 1.5 Bit Per Stage Pipelined Converter

Adding a second comparator to each stage of a pipelined converter results in the 1.5-bit[1] stage architecture shown in Fig. 17.20 [Lewis, 1992]. Each stage effectively performs a 3-level quantization of its input,

$$V_{i,x} = (b_{i,0} - 0.5)\frac{V_{ref}}{4} + (b_{i,1} - 0.5)\frac{V_{ref}}{4} = (b_{i,0} + b_{i,1} - 1)\frac{V_{ref}}{4} \tag{17.41}$$

This estimate is then subtracted from the input and scaled by $2\times$ to form the residue V_{i+1}. Equation (17.41) shows that both $b_{i,0}$ and $b_{i,1}$ contribute $\pm V_{ref}/4$ to the residue and therefore they are weighted equally in determining the converter's digital output. The final stage of is typically a simple two-bit A/D, since no residue is needed. Assuming the final stage has a two-bit output whose MSB and LSB are $b_{N-1,1}$ and $b_{N-1,0}$ respectively, the final quantized signal is

$$\hat{V}_{in} = \frac{V_{ref}}{2}\left(\sum_{i=1}^{N-2}(b_{i,0} + b_{i,1} - 1)2^{-i} + b_{N-1,1}2^{N-2} + b_{N-1,0}2^{N-1}\right) \tag{17.42}$$

The output bits of all stages are converted into the final output code by addition with the proper binary weighting, as in (17.42). Of course, prior to addition, the outputs of all stages must be delayed by varying length shift registers to realign the bits in time, as shown in Fig. 17.21.

1. Strictly speaking, because each stage produces one of three digital output codes (00, 01, or 11), they are actually $\log_2(3) = 1.585$-bit stages, but in practice they are always referred to simply as 1.5-bit stages.

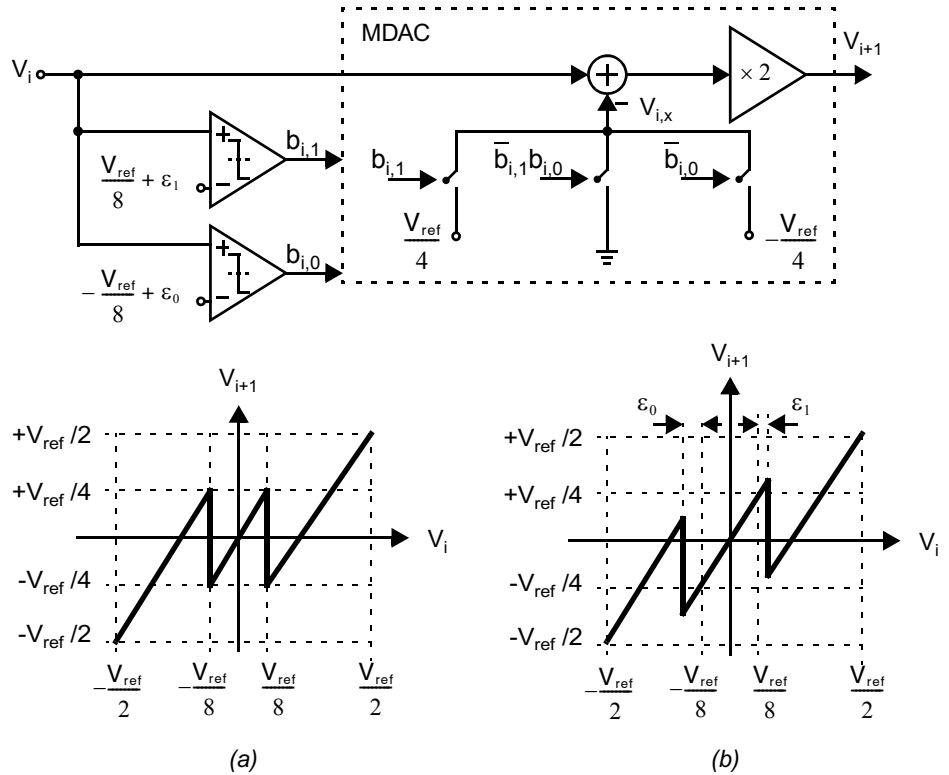


Fig. 17.20 A 1.5-bit pipelined converter stage that accepts inputs in the range $-V_{ref}/2 < V_i < V_{ref}/2$. (*a*) The ideal input–output relationship. (*b*) The input–output relationship in the presence of comparator offset.

Although the 1.5 bit per stage architecture entails some digital decoding logic, its redundancy makes the converter robust to comparator offsets. The impact of random comparator offsets $\varepsilon_0, \varepsilon_1$ on the pipeline stage's input–output relationship is shown in Fig. 17.20(*b*). Intuitively the offsets will, for some input levels, introduce an error of $\pm V_{ref}/4$ into $V_{i,x}$ which is reflected in the values of $b_{i,0}$ and $b_{i,1}$. They also result in an error of $\mp V_{ref}/2$ in V_{i+1}, which is captured by the output bits of subsequent stages in the pipeline. When all bits are combined by the decoder (17.42) the errors cancel, as long as $|\varepsilon_{0,1}| < V_{ref}/4$ so that the residue remains within the range $\pm V_{ref}/2$. This argument is inductive; comparator offsets in all 1.5 bit stages are cancelled by subsequent stages. Only offsets in the final stage are left uncanceled. Fortunately, such offsets influence only the converter's 2 LSBs, and are therefore rarely a significant limitation in the design of pipelined converters.

Key Point: *Introducing redundancy between successive stages in a pipelined converter permits simple digital correction of any offset in the comparators of each stage.*

Errors introduced in the MDAC generally limit the performance of pipelined converters. For example, mismatch resulting in errors the DAC voltages $\pm V_{ref}/4$, inaccuracy in the $2\times$ gain, and thermal noise in the MDAC appear in the residue V_{i+1} but not the comparator outputs $b_{i,0}$ and $b_{i,1}$, and are therefore not cancelled. The impact of such errors is greatest at the start of the pipeline, in stage 1 where they influence the converter's most significant bits. Therefore, the first stage is normally designed to be the larger and consume more power than subsequent stages in order to minimize its mismatch and noise.

Key Point: *Pipelined A/D converter performance is limited by nonlinearities and noise in the MDAC, and inaccuracy in the interstage gain, none of which are cancelled by simple digital correction.*

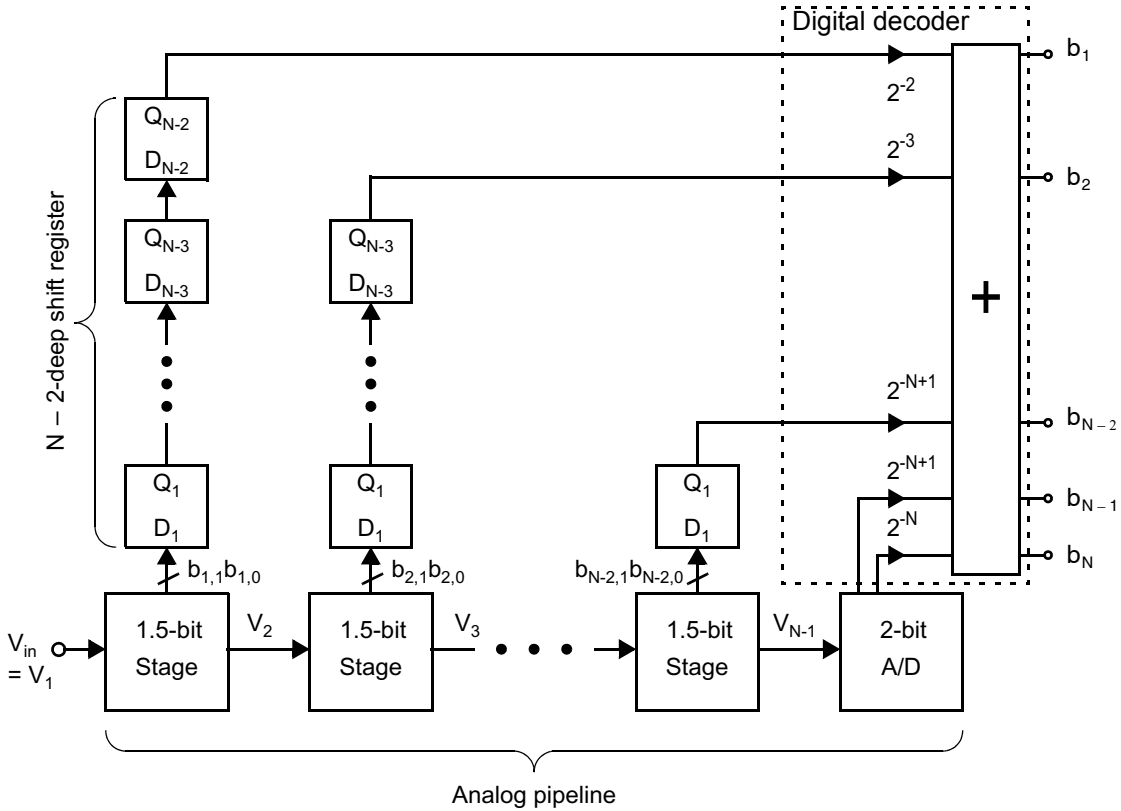


Fig. 17.21 A 1.5-bit-per-stage pipelined A/D converter.

EXAMPLE 17.8

Consider the operation of a 3-bit, 1.5-bit-per-stage pipelined converter. Take $V_{ref} = 1$ V and $V_{in} = 150$ mV. Find the residue voltage, V_2, the digital codes at the outputs of each stage, and the final digital output after decoding. Then, repeat the exercise assuming the upper comparator of stage 1 has an offset of 30 mV.

Solution

In the absence of comparator offset, the digital outputs of stage 1 will be $b_{1,1} = 1$ and $b_{1,0} = 1$ and the DAC output is $V_{1,x} = V_{ref}/4 = 250$ mV and a residue of

$$V_2 = 2(V_1 - V_{1x}) = -200 \text{ mV}$$

This is input to the final stage, a 2-bit A/D having a full-scale range of $\pm V_{ref}/2 = \pm 500$ mV. Its binary output is $(01)_2$. Hence, the digital code at the output of the decoder is computed as follows:

$b_{1,1}$:	1
$b_{1,0}$:	1
$b_{2,1}b_{2,0}$:	01
$b_1b_2b_3$:	101

Next, consider what happens if the upper comparator of the first stage has a 30-mV offset, sufficient to cause a error in its output: $b_{1,1} = 0$. As a result, $V_{1,x} = 0$ and the residue becomes

$$V_2 = 300 \text{ mV}$$

The binary output of the final-stage 2-bit A/D therefore becomes $(11)_2$. The following addition is performed by the decoder:

$b_{1,1}$:	0
$b_{1,0}$:	1
$b_{2,1}b_{2,0}$:	11
$b_1b_2b_3$:	101

The error in the first stage is captured by the following stage resulting in the same decoder output as in the ideal case.

17.4.3 Pipelined Converter Circuits

Key Point: *Switching part of the sampling capacitor into the feedback network provides the interstage gain with a higher feedback factor, and hence higher bandwidth, than a conventional switched-capacitor gain circuit.*

The most common implementation of the 1.5 bit stage is shown in Fig. 17.22 [Sutarja, 1988], although fully-differential implementations are used in most high-performance applications. In the first clock phase, the input V_i is sampled on the total capacitance $2C$, Fig. 17.22(*a*). Then in the second clock phase, Fig. 17.22(*b*), one-half of the total capacitance is switched to become the feedback capacitance. Reusing the sampling capacitance in this way realizes a gain of 2 with a feedback factor of $\beta = 1/2$ in Fig. 17.22(*b*). In a conventional switched-capacitor gain stage, a separate feedback capacitance C is used in addition to the $2C$ input sampling capacitance resulting in the lower feedback factor $\beta = 1/3$ and, hence, lower closed-loop bandwidth and a lower maximum clock frequency.

The other one-half of the sampling capacitance is split into two $C/2$ capacitors and reused as a charge redistribution D/A converter during the second clock phase, Fig. 17.22(*b*). The 1.5-bit code $b_{i,1}$, $b_{i,0}$ serves as the digital input to the D/A and introduces a net change of $\pm V_{ref}/2$ or 0 at the output, V_{i+1}.

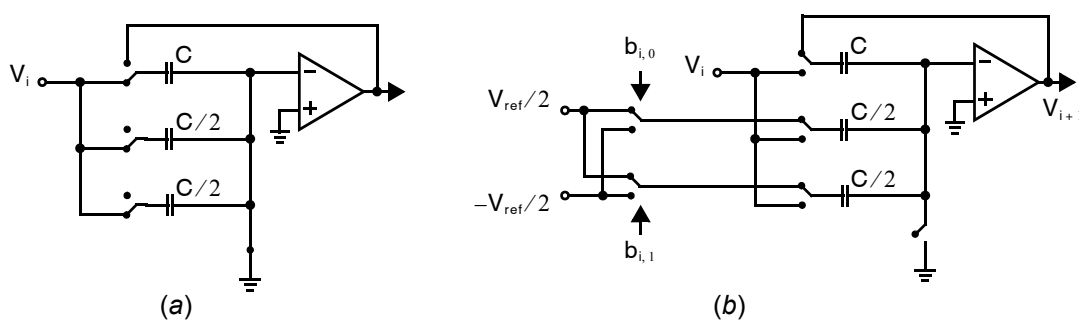


Fig. 17.22 MDAC for the 1.5 bit per stage pipelined converter during two clock phases: (*a*) input is sampled onto all capacitors; (*b*) the input is transferred to the output with a gain of two and the comparator output bits $b_{i,0}$ and $b_{i,1}$ determine the subtracted signal.

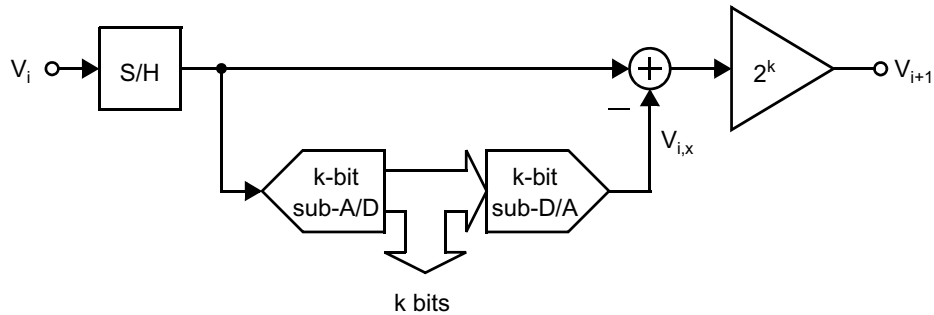


Fig. 17.23 A general k-bit stage in a pipelined A/D converter.

17.4.4 Generalized k-Bit-Per-Stage Pipelined Converters

More than one bit can be resolved per stage by increasing the interstage gain and number of comparators per stage. The signal flow is similar to that illustrated in Fig. 17.13 for a 2-bit per iteration successive approximation converter. A generalized k-bit stage is shown in Fig. 17.23. Nonlinearities in the k-bit sub-A/D converter can be digitally corrected by adding additional comparators, similar to the 1.5 bit per stage architecture [Lewis, 1992]. The advantages of the k-bit-per-stage architecture are that fewer stages are needed, which can reduce area and/or power consumption compared with the 1.5 bit per stage. Also, for a given resolution, there are fewer stages for the input signal to traverse through the pipeline, so there are fewer clock cycles of latency in each conversion.

The circuit shown in Fig. 17.22 can be modified to accommodate k bits per stage with $k \geq 2$ by dividing the input sampling capacitor into smaller unit-sized devices. The required interstage gain is greater than 2, so a smaller fraction of the sampling capacitance is switched into the feedback network in the second clock phase. Hence, the feedback factor β is reduced and closed loop bandwidth is decreased compared with the 1.5-bit stage (assuming the same opamp is used).

Key Point: *Multiple bits can be resolved in each stage, with digital correction, in order to reduce the number of opamps in a pipelined converter and decrease the number of clock cycles required for each conversion.*

The major limitations on converter accuracy remains the accuracy of the interstage gain, linearity of the sub-D/A converter, and noise in the MDAC circuit. Again, accuracy requirements are most stringent in the first stage where errors influence the most significant bits.

17.5 FLASH CONVERTERS

Flash converters are the standard approach for realizing very-high-speed converters. The input signal in a flash converter is fed to 2^N comparators in parallel, as shown in Fig. 17.24. Each comparator is also connected to a different node of a resistor string. Any comparator connected to a resistor string node where V_{ri} is larger than V_{in} will have a 1 output while those connected to nodes with V_{ri} less than V_{in} will have 0 outputs. Such an output code word is commonly referred to as a thermometer code since it looks quite similar to the mercury bar in a thermometer. Note that the top and bottom resistors in the resistor string have been chosen to create a 0.5 LSB offset in an A/D converter.

Key Point: *Flash converters are generally capable of the fastest conversion speeds of any A/D architecture. However, their power consumption grows dramatically with the required resolution so they are usually reserved for converters where high speed and low or modest resolution is sought.*

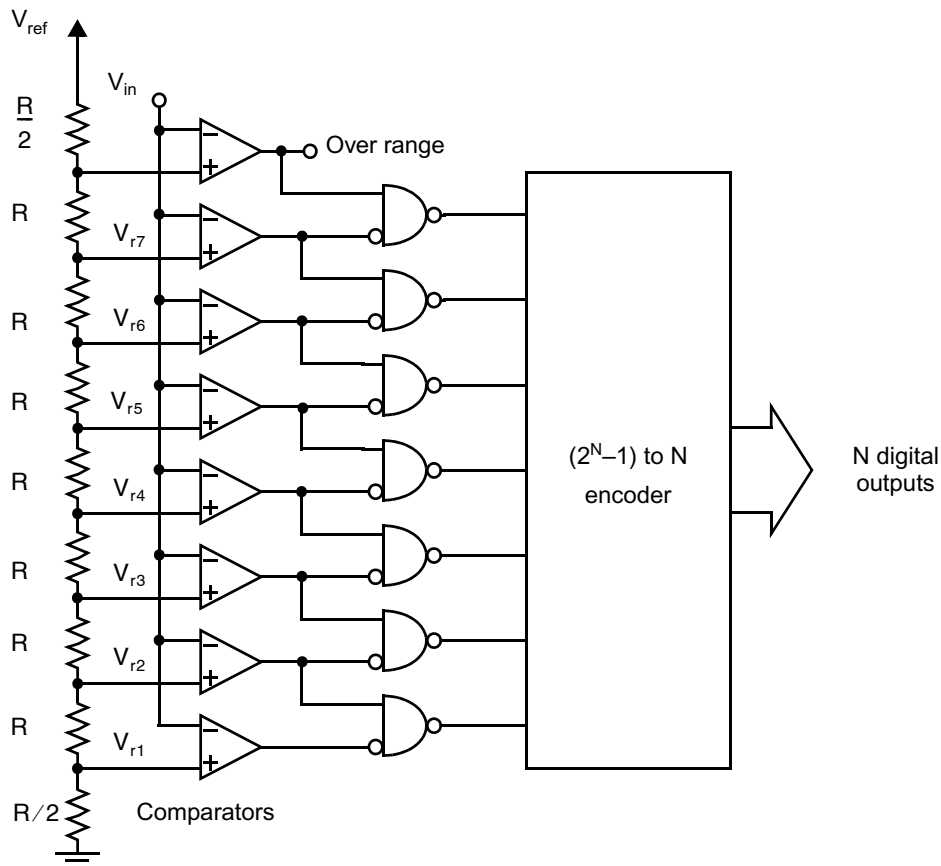


Fig. 17.24 A 3-bit flash A/D converter.

The NAND gate that has a 0 input connected to its inverting input and a 1 input connected to its noninverting input detects the transition of the comparator outputs from 1s to 0s, and will have a 0 output. All other NAND-gate outputs will be 1, resulting in simpler encoding. It also allows for error detection by checking for more than one 0 output, which occurs during a bubble error (see the next subsection) and, perhaps, error correction.

Flash A/Ds are fast but the number of comparators grows exponentially with the resolution N, so they typically take up a large area and are very power hungry, even for modest N—especially when they are clocked fast. One way to realize a small clocked CMOS comparator is by using a CMOS inverter, as shown in Fig. 17.25 [Dingwall, 1979].

When ϕ is high, the inverter is set to its bistable operating point, where its input voltage equals its output voltage (i.e., its threshold voltage). Normally with an odd number of inverters, a ring oscillator is formed; however, in the case of a single CMOS inverter, the inverter operates as a single stage opamp with only one pole (no nondominant poles), so stability is guaranteed. With this inverter set to its threshold voltage, the other side of C is charged to V_{ri}. When ϕ goes low, the inverter is free to fall either high or low depending on its input voltage. At the same time, the other side of C is pulled to the input voltage, V_{in}. Since the inverter side of the capacitor is floating, C must keep its original charge, and therefore the inverter's input will change by the voltage difference between V_{ri} and V_{in}. Since the inverter's input was at its bistable point, the difference between V_{ri} and V_{in} will determine which direction the inverter's output will fall. However, it should be mentioned that this simple comparator suffers from poor power supply rejection, which is often a critical design specification in fast converters. Using fully differential inverters helps alleviate this shortcoming.

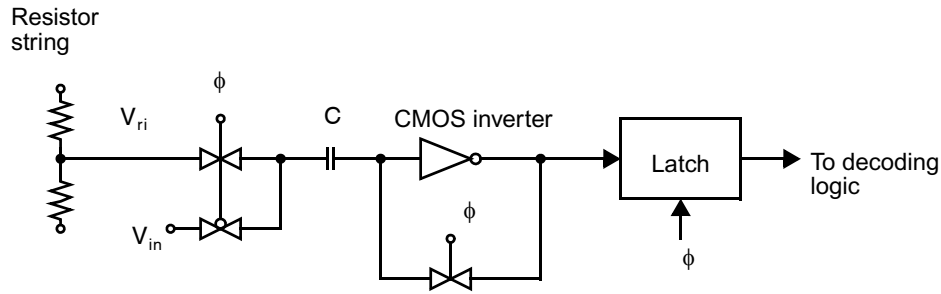


Fig. 17.25 A clocked CMOS comparator.

17.5.1 Issues in Designing Flash A/D Converters

We discuss here some important design issues that should be addressed when building high-speed flash A/D converters.

Input Capacitive Loading The large number of comparators connected to V_{in} results in a large parasitic load at the node V_{in}. Such a large capacitive load often limits the speed of the flash converter and usually requires a strong and power-hungry buffer to drive V_{in}. We shall see that this large capacitive loading can be reduced by going to an interpolating architecture.

Resistor-String Bowing Any input currents to the comparators cause errors in the voltages of the nodes of the resistor string. These errors usually necessitate the bias current in the resistor string being two orders of magnitude greater than the input currents of the comparators. This is particularly significant if bipolar comparators are used. The errors are greatest at the center node of the resistor string and thus considerable improvement can be obtained by using additional circuitry to force the center tap voltage to be correct.

Comparator Latch-to-Track Delay Another consideration that is often overlooked is the time it takes a comparator latch to come from latch mode to track mode when a small input signal of the opposite polarity from the previous period is present. This time can be minimized by keeping the time constants of the internal nodes of the latch as small as possible. This is sometimes achieved by keeping the gain of the latches small, perhaps only two to four. In many cases, the differential internal nodes might be shorted together temporarily as a reset just after latch time.

Signal and/or Clock Delay Even very small differences in the arrival of clock or input signals at the different comparators can cause errors. To see this, consider a 250-MHz, 1-V peak-input sinusoid. This signal has a maximum slope of 1570 V/μs at the zero crossing. If this signal is being encoded by an 8-bit A/D converter with $V_{ref} = 2$ V (i.e., the sinusoid covers the whole range), then it would only take 5 ps to change through 1 LSB. This time is roughly about the same time it takes a signal to propagate 500 μm in metal interconnect. If there is clock skew between comparators greater than this, the converter will have more than 1 LSB error. One means of easing this problem is to precede the converter by a sample-and-hold circuit. However, high-speed sample-and-hold circuits can be more difficult to realize than the flash converter itself. In addition, the clock and V_{in} should be routed together with the delays matched [Gendai, 1991]. It should also be noted that the delay differences may not be caused just by routing differences, but could also be caused by different capacitive loads, or by phase differences between the comparator preamplifiers at high frequencies.

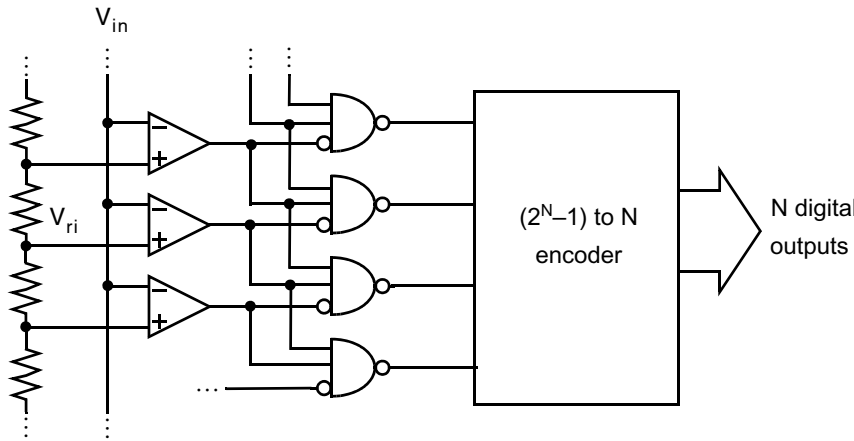


Fig. 17.26 Using three-input NAND gates to remove single bubble errors.

Substrate and Power-Supply Noise For $V_{ref} = 1$ V and an 6-bit converter, only 15.6 mV of noise injection would cause a 1 LSB error. On an integrated circuit having a clock signal in the hundreds or even thousands of MHz, it is difficult to keep power-supply noise below 100 mV. This power-supply noise can easily couple through the circuitry or substrate, resulting in errors. To minimize this problem, the clocks must be shielded from the substrate and from analog circuitry. Also, running differential clocks closely together will help prevent the signals from being coupled into the substrate or to critical circuit nodes through the air. Also, analog power supplies should be separated from digital power supplies. Specifically, it is advisable to have analog power supplied to the comparator preamps while using digital power for the latch stages. On-chip power-supply bypass capacitors are a necessity. It is also necessary to make sure the power-supply bypass circuitry doesn't resonate with the circuit packaging parasitics. Small resistors in series with the by-pass capacitors can help dampen any such resonance.

Bubble Error Removal The outputs of the comparators should be a thermometer code with a single transition. However, sometimes a lone 1 will occur within the string of 0s (or a 0 within the string of 1s) due to comparator metastability, noise, cross talk, limited bandwidth, etc. These bubbles usually occur near the transition point of the thermometer code. Fortunately, these bubbles can usually be removed with little extra complexity by replacing the two-input NAND gates shown in Fig. 17.24 with three-input NAND gates, as shown as shown in Fig. 17.26 [Steyaert, 1993]. With this modification, there must now be two 1s immediately above a 0 in determining the transition point in the thermometer code. However, this circuit will not eliminate the problem of a stray 0 being two places away from the transition point, which may cause a large decoding error. Another digital approach for reducing the effect of bubble errors is to allow bubble errors in the lower 2 LSBs but have the remaining MSBs determined by looking for transitions between every *fourth* comparator [Gendai, 1991]. With this approach, bubble errors that occur within four places of the transition point do not cause any large errors. An alternate approach to reduce the effect of distant bubble errors is to create two encoders (one AND type and one OR type) rather than a single encoder [Ito, 1994]. When an unexpected output pattern occurs at the NAND outputs, the errors in two different encoders tend to be equal in magnitude but opposite in sign. Thus, the final output is taken as the average of the two encoder outputs, which is performed by adding the two outputs and dropping the LSB (to divide by two).

An alternative method to remove bubble errors that does not increase power dissipation is shown in Fig. 17.27. Here, extra transistors have been added to the inputs of the slave latches, which are driven by the comparator master latches [van Valburg, 1992]. These extra transistors make the value stored in a slave latch not just a function of its master latch, but also a function of the two adjacent master latches. If a bubble occurs, the outputs from the two adjacent master latches are the same, but different from the center master latch. In this case,

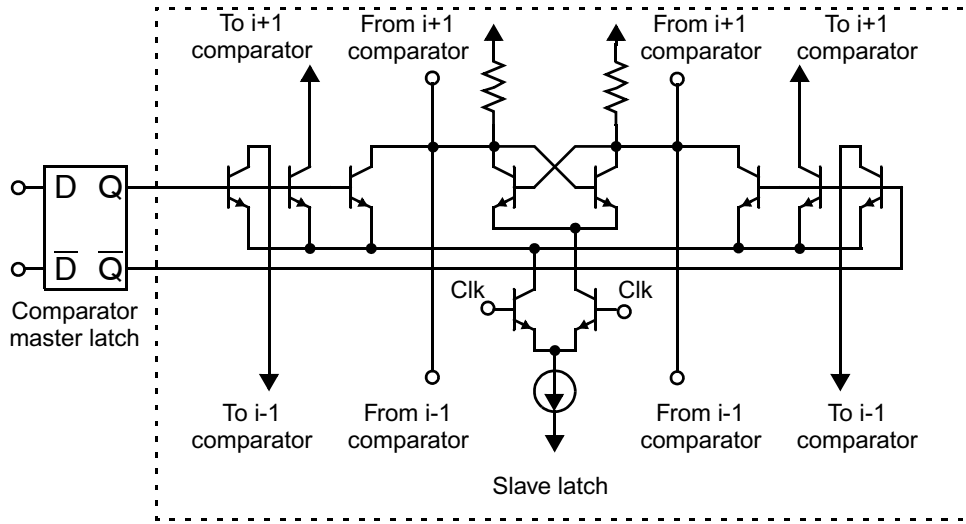


Fig. 17.27 Bubble-error voting circuit that does not increase the power dissipation.

the values from the adjacent master latches overrule the center master latch. Note that with this approach, the power dissipation is not increased because the added transistors make use of existing current in the slave latch. Alternatively, this same voting scheme can be implemented entirely in digital form.

Flashback An additional source of error is flashback**.** Flashback is caused by clocked comparators, which are almost always used. When clocked comparators are switched from track to latch mode, or vice versa, there is major charge glitch at the inputs to the latch. If there is no preamplifier, this will cause major errors due to the unmatched impedances at the comparator inputs (one input goes to the resistor string—the other to the input signal). To minimize this effect, most modern comparators have one or two stages of continuous-time buffering and/or preamplification. For example, a bipolar comparator is shown in Fig. 17.28. Notice that it has a buffer, a low-gain preamp, and another buffer before the track-and-latch circuitry. Also notice that in the positive feedback latch the feedback is taken from the emitter-follower outputs, which minimizes the capacitances of the internal nodes of the latch. A CMOS comparator may be used with the same topology, although the source-follower buffers are usually omitted.

Another technique sometimes used to minimize the effects of flashback is to match the input impedances as much as is possible. For example, it is possible to implement a second matched resistor string, with the nodes of the comparators that were originally connected to V_{in} now being connected to it, and the end nodes of the string connected together to V_{in}. This approach matches impedances and also minimizes the resistor-string bowing due to the comparator input currents. Unfortunately, it does result in different delays for V_{in} reaching the various comparators, and unless these are matched to the routing of the clock signals, these different delays may not be tolerable.

17.6 TWO-STEP A/D CONVERTERS

Two-step (or subranging) converters are used for high-speed medium-accuracy A/D converters. They offer several advantages over their flash counterparts. Specifically, two-step converters require less silicon area, dissipate less power, have less capacitive loading, and the voltages the comparators need to resolve are less stringent than for flash equivalents. The throughput of two-step converters approaches that of flash converters, although they do have a larger latency.

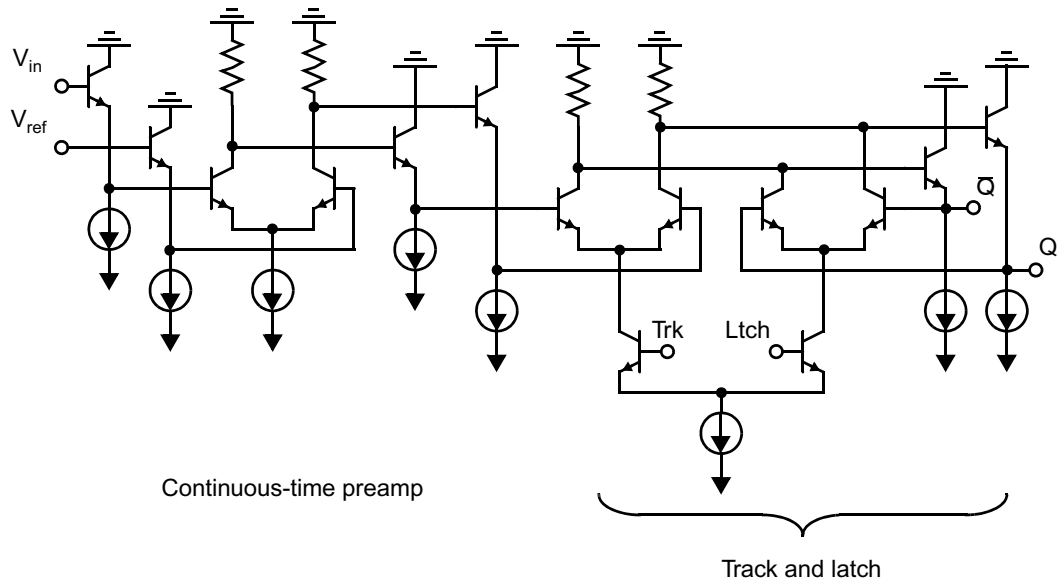


Fig. 17.28 Clocked comparator with a preamplifier to reduce flashback.

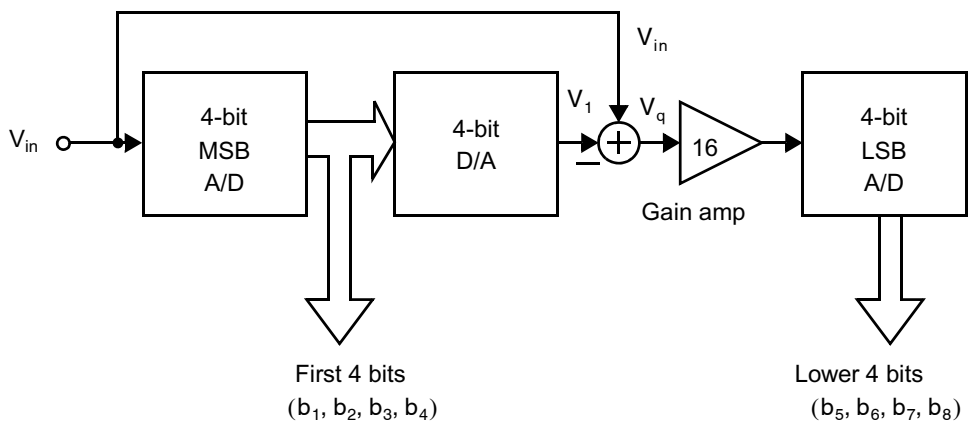


Fig. 17.29 An 8-bit two-step A/D converter.

The block diagram for a two-step converter is shown in Fig. 17.29. In fact, a two-step converter may be thought of as a special case of a pipelined converter with only two pipeline stages. The example of Fig. 17.29 shows two 4 bit stages with no redundancy resulting in a 8-bit output. The 4-bit MSB A/D determines the first four MSBs. To determine the remaining LSBs, the quantization error (residue) is found by reconverting the 4-bit digital signal to an analog value using the 4-bit D/A and subtracting that value from the input signal. To ease the requirements in the circuitry for finding the remaining LSBs, the quantization error is first multiplied by 16 using the gain amplifier, and the LSBs are determined using the 4-bit LSB A/D. With this approach, rather than requiring 256 comparators as in an 8-bit flash converter, only 32 comparators are required for a two-step A/D converter. However, this straightforward approach would require all components to be at least 8-bit accurate. To significantly ease the accuracy requirements of the 4-bit MSB A/D converter, digital error correction is commonly used.

17.6.1 Two-Step Converter with Digital Error Correction

The block diagram for a two-step converter with digital error correction is shown in Fig. 17.30. Its operation is directly analogous to that of the 1.5 bit per stage pipeline converter. The ranges of the 4-bit and 5-bit stages overlap, and hence inaccuracies in the first A/D converter can be cancelled by the second. To see how this correction works and why a second-stage 5-bit converter is needed (rather than 4-bit), consider the quantization error that occurs in an ideal converter. Defining $V_{LSB} = V_{ref}/2^8$ (i.e., always relative to 8-bit accuracy), we have for an ideal 8-bit converter,

> ***Key Point:*** *Two-step converters may be considered a special case of pipelined converters where only 2 pipelined stages are needed. As with pipelined converters, digital error correction greatly relaxes the requirements on the first stage.*

$$V_{ref}B_{out} = V_{in} + V_q \qquad \text{where} \quad -\frac{1}{2}V_{LSB} < V_q < \frac{1}{2}V_{LSB} \tag{17.43}$$

However, for a nonideal 8-bit converter with an absolute accuracy of 0.5 LSB, we have

$$V_{ref}B_{out} = V_{in} + V_q \qquad \text{where} \quad -V_{LSB} < V_q < V_{LSB} \tag{17.44}$$

In other words, the maximum quantization signal is now twice that of the ideal case.

Similarly, for an ideal 4-bit A/D converter, we have (keeping the 8-bit definition of V_{LSB}),

$$V_{ref}B_{out} = V_{in} + V_q \qquad \text{where} \quad -8V_{LSB} < V_q < 8V_{LSB} \tag{17.45}$$

Thus in the ideal case, the value of V_q can be determined (to 8-bit accuracy) using a 4-bit A/D converter since V_q must be within $16V_{LSB}$. However, for the nonideal case where the 4-bit MSB flash converter has an absolute accuracy of $8V_{LSB}$, the quantization error, V_q, is now bounded within $32V_{LSB}$. Thus, in the case of a non-ideal 4-bit MSB converter, a 5-bit LSB converter must be used; otherwise V_q may go out of range. Note that the gain amplifier of 8 is used to amplify the quantization error back to maximum signal levels to ease the requirements of the 5-bit LSB converter. Finally, to determine V_{in}, we see that the digital value of V_q has been found to within

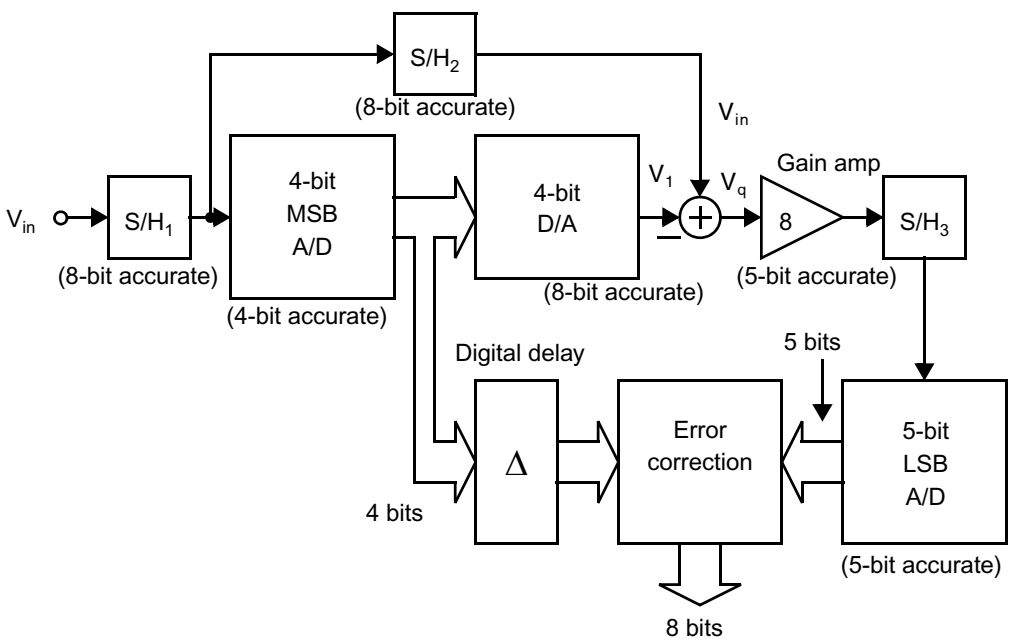


Fig. 17.30 An 8-bit two-step A/D converter with digital error correction.

$0.5V_{LSB}$ and the digital value of V_1 is known to the same accuracy since we assumed the D/A converter to be 8-bit accurate (and the digital word applied to the D/A converter is known). Therefore, we find V_{in} from the relation

$$V_{in} - V_1 = V_q \tag{17.46}$$

Specifically, V_{in} is found by properly combining the digital equivalents of V_1 and V_q.

In summary, the MSB A/D need only be accurate to $1/2^4 = 1/16$. The only components that need 0.5 LSB accuracy at the 8-bit level are S/Hs 1 and 2, the D/A, and the subtraction circuit. For a more detailed treatment of a two-step A/D converter, the reader is referred to a 10-bit 75 MHz implementation (with integrated S/H), described in [Petschacher, 1990].

Besides the difficulty in realizing the S/H circuits, another major limitation is the difficulty of designing a high-speed, accurate (to 0.5 LSB at the 5-bit level) gain amplifier. In fact, due to difficulties in realizing high-speed circuits with gain, often fewer bits are determined in the first stage of a two-step converter to reduce the amplification required. For example, a two-step converter where the first stage resolves only 1 bit is described in [Verbruggen, 2009] achieving very low power.

EXAMPLE 17.9

For the two-step 8-bit A/D converter shown in Fig. 17.30, what is the maximum voltage range at V_q when the converter's full-scale input is ± 2.5 V in the case where the 4-bit MSB A/D converter is (a) 8-bit accurate and (b) 4-bit accurate? Assume all other components are ideal.

Solution

With a full-scale peak-to-peak input voltage of 5 V applied to an 8-bit A/D converter, we have

$$V_{LSB} = \frac{5}{2^8} = 19.5 \text{ mV} \tag{17.47}$$

For an ideal 4-bit A/D converter, we have from (17.45) the maximum voltage range of V_q is $16V_{LSB}$, or equivalently, 312 mV.

(a) If we go to the trouble to make the 4-bit A/D converter have an absolute accuracy of $0.5V_{LSB}$ (i.e., 8-bit accurate), then V_q becomes bounded between $\pm 8.5V_{LSB}$. In other words, the maximum range of V_q would now be $17V_{LSB}$, or equivalently, 332 mV. Note that the input range of the LSB converter is 8×332 mV = 2.7 V (a little more than half of the input range of the overall converter), so more gain could be used in this case.

(b) In the case of a 4-bit accurate MSB converter, V_q is bounded between $\pm 16V_{LSB}$, implying that the maximum range of V_q is $32\ V_{LSB}$, or equivalently, 618 mV. After the gain of 8, the input range of the LSB converter becomes 8×618 mV = 4.9 V (the same as the input range of the overall converter).

17.7 INTERPOLATING A/D CONVERTERS

Key Point: *The large capacitive input presented by flash A/D converters can be reduced by using fewer input amplifiers and interpolating between them to create the missing intermediate values.*

Interpolating converters make use of input amplifiers, as shown in Fig. 17.31. These input amplifiers behave as linear amplifiers near their threshold voltages but are allowed to saturate once their differential inputs become moderately large. As a result, noncritical latches need only determine the sign of the amplifier outputs since the differences between the input signal and threshold voltages have been amplified. Also, the number of input amplifiers attached to V_{in} is significantly reduced by interpolating between adjacent outputs of these amplifiers. While this approach is often combined with a "folding" architecture [van de Grift, 1987; van Valburg, 1992], the interpolating architecture has also been used quite successfully by itself [Goodenough, 1989; Steyaert, 1993].

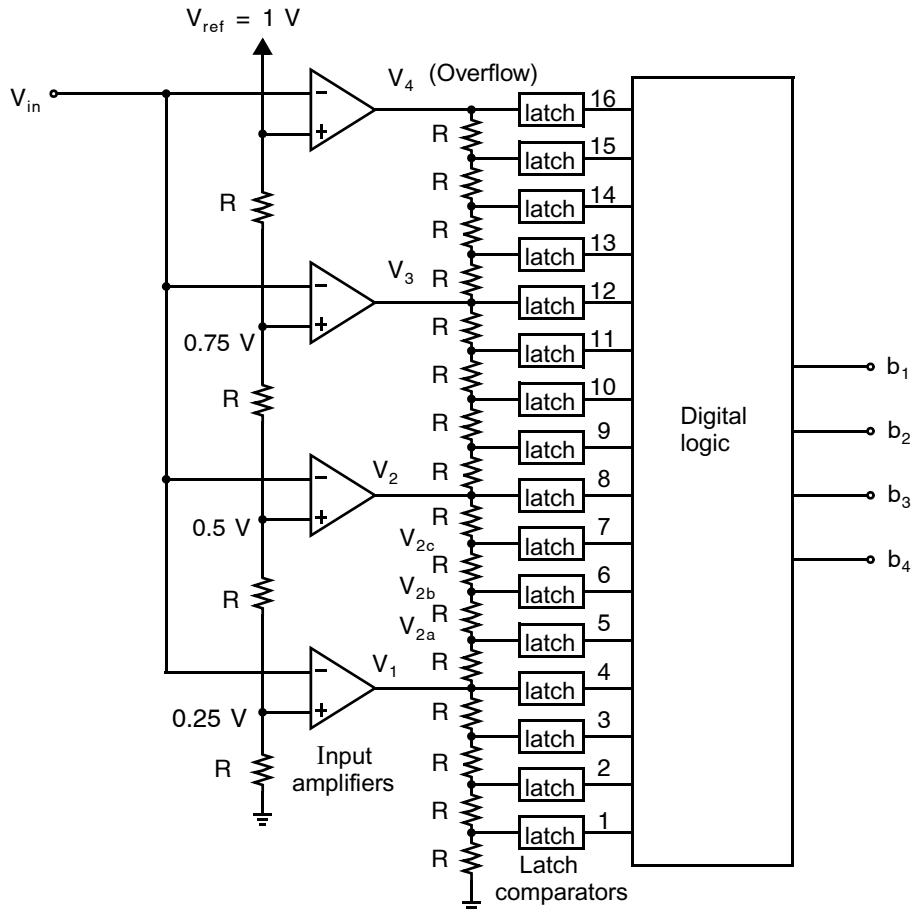


Fig. 17.31 A 4-bit interpolating A/D converter (interpolating factor of 4).

To further understand this interpolation approach, some possible signals for the input-amplifier outputs, V_1 and V_2, as well as their interpolated values are shown in Fig. 17.32. As can be seen in the figure, the logic levels are assumed to be zero and five volts, with the input comparators having a maximum gain of about –10. Also, here the latch threshold is near the midpoint of the two logic levels (or about 2.5 volts). As V_{in} increases, the latch for V_1 is first triggered, followed by V_{2a} and so on until V_2. As a result, more reference levels have been created between V_1 and V_2. It should be noted here that for good linearity, the interpolated signals need only cross the latch threshold at the correct points, while the rest of the interpolated signals responses are of secondary importance. One way to create such correct crossing points is to ensure that V_1 and V_2 are linear between their own thresholds. In Fig. 17.32, this linear region corresponds to $0.25 < V_{in} < 0.5$.

For fast operation, it is important that the delays to each of the latches are made to equal each other as much as possible. Since the latch comparators have similar input capacitances associated with them, the delays can be made nearly equal by adding extra series resistors, as shown in Fig. 17.33. These series resistors equalize the impedances seen by each latch comparator looking back into the resistive string, assuming the input-amplifier outputs are low impedance [van de Plassche, 1988].

As mentioned earlier, the main benefit of an interpolating architecture is the reduction in the number of differential pairs attached to the input signal, V_{in}. Such a reduction results in a lower input capacitance (which is quite

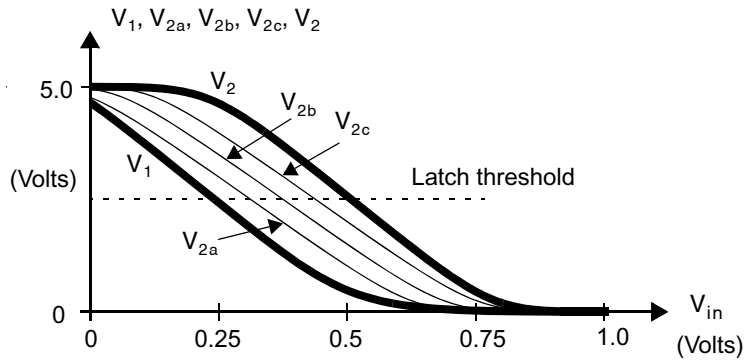


Fig. 17.32 Possible transfer responses for the input-comparator output signals, V_1 and V_2, and their interpolated signals.

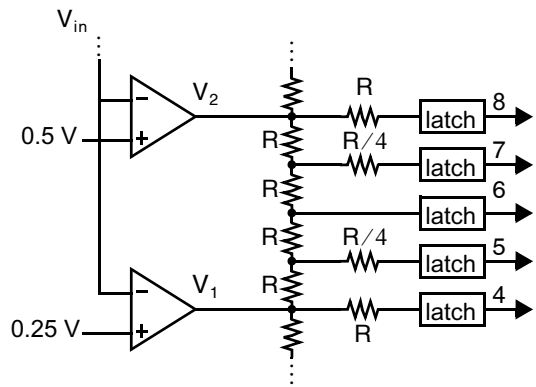


Fig. 17.33 Adding series resistors to equalize delay times to the latch comparators.

high for a flash converter), a slightly reduced power dissipation, and a lower number of accurate reference voltages that need to be created. Finally, it should be mentioned that circuit techniques other than resistive strings can be used to realize this interpolative approach. In [Steyaert, 1993], current mirrors were used to interpolate eight times between comparators. In another implementation, two stages of interpolation using capacitors were realized [Kusumoto, 1993].

EXAMPLE 17.10

Using current mirrors, show how one can interpolate two current outputs, I_1 and I_2, by three. What reduction in input capacitance of the converter would be expected over a traditional flash architecture?

Solution

If interpolating by three, it is desired to create two new currents, I_{2a}, I_{2b} such that

$$I_{2a} = \frac{2}{3}I_1 + \frac{1}{3}I_2 \tag{17.48}$$

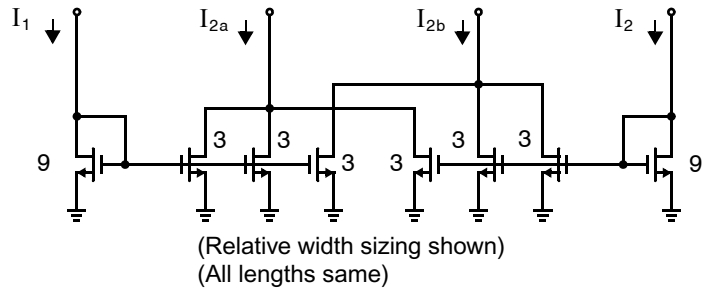


Fig. 17.34 Interpolating by three between two current outputs.

$$I_{2b} = \frac{1}{3}I_1 + \frac{2}{3}I_2 \tag{17.49}$$

These output currents can be realized as shown in Fig. 17.34. These four currents can be converted back to voltages to send to the latches. Alternatively, the currents can be directly sent to latches which make use of current inputs.

Since we are interpolating by three here, this converter would require one-third the number of input amplifiers in relation to a traditional flash converter. Thus, the input capacitance for this interpolated approach would be one-third of that for a flash.

17.8 FOLDING A/D CONVERTERS

Key Point: *A folding architecture requires fewer latches than a flash converter. It does so similar to a two-step converter, but uses analog preprocessing instead of an accurate D/A converter and gain amplifier.*

We just saw that the number of input amplifiers can be reduced through the use of an interpolating architecture. However, the number of latch comparators remains at 2^N for an N-bit converter. This large number of latch comparators can be significantly reduced through the use of a *folding architecture*. A folding A/D converter is similar in operation to a two-step (or subranging) converter in that a group of LSBs are found separately from a group of MSBs. However, whereas a two-step converter requires an accurate D/A converter, a folding converter determines the LSB set more directly through the use of analog preprocessing while the MSB set is determined at the same time.

As an example, consider the 4-bit folding converter shown in Fig. 17.35. Defining the *folding rate* to be the number of output transitions for a single folding block as V_{in} is swept over its input range, we see that the folding rate here is four. This folding rate determines how many bits are required in the MSB converter. The operation of this converter is as follows. The MSB converter determines whether the input signal, V_{in}, is in one of four voltage regions (i.e., between 0 and 1/4, 1/4 and 1/2, 1/2 and 3/4, or 3/4 and 1). Although the MSB converter is shown separately, these bits are usually determined by combining appropriate signals within the folding blocks. To determine the 2 LSBs, V_1 to V_4 produce a thermometer code for each of the four MSB regions. Note, however, that the four LSB latches are also used for different MSB regions and the thermometer code is inverted when V_{in} is between either 1/4 and 1/2 or 3/4 and 1. For example, as V_{in} increases from 0 to 1/4, the thermometer code changes as 0000, 0001, 0011, 0111, 1111. However, as V_{in} continues to increase to 1/2, the code changes as 1110, 1100, 1000, 0000. Also, note that latch comparators can be used for the LSB set since the transitions are amplified by the folding blocks. In summary, folding reduces the number of latch comparators needed as compared to a flash

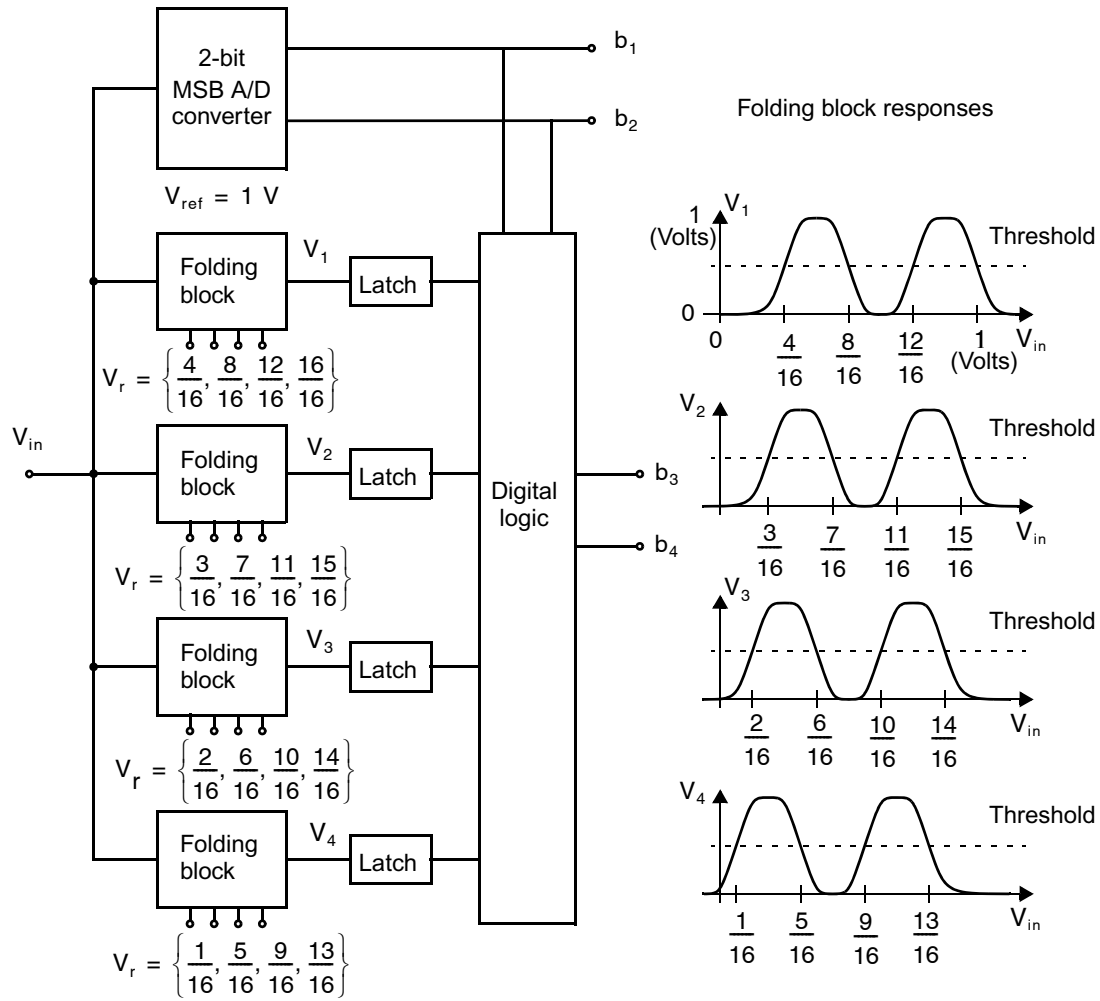


Fig. 17.35 A 4-bit folding A/D converter with a folding rate of four. (The MSB converter would usually be realized by combining some folding block signals.)

converter. For example, in a flash 4-bit converter, 16 latches would be required whereas only eight are needed in the 4-bit folding example of Fig. 17.35. Specifically, four latches are used for the MSB converter and the other four are shown explicitly. In general, the savings can be greater (see Problem 17.29).

The folding blocks can be realized using cross-coupled differential pairs, as seen in the simplified bipolar circuit shown in Fig. 17.36. Here, four sets of differential-pair transistors are connected in such a way as to realize the input-output response shown in Fig. 17.36(*b*). The output signal V_{out} is related to the voltages V_a and V_b in an "or" type fashion. In other words, V_{out} is low only if both V_a and V_b are low; otherwise, V_{out} is high. With regard to the behavior of V_a and V_b, note that V_b remains low whenever V_{in} is less than V_{r3} or greater than V_{r4}. However, V_a remains low whenever V_{in} is greater than V_{r2} or less than V_{r1}. Also, the cross coupling of adjacent differential pairs causes V_a to go high when V_{in} is between V_{r1} and V_{r2}, while V_b goes high when V_{in} is between V_{r3} and V_{r4}. Such behaviors for V_a and V_b give rise to the folding output for V_{out}, as shown.

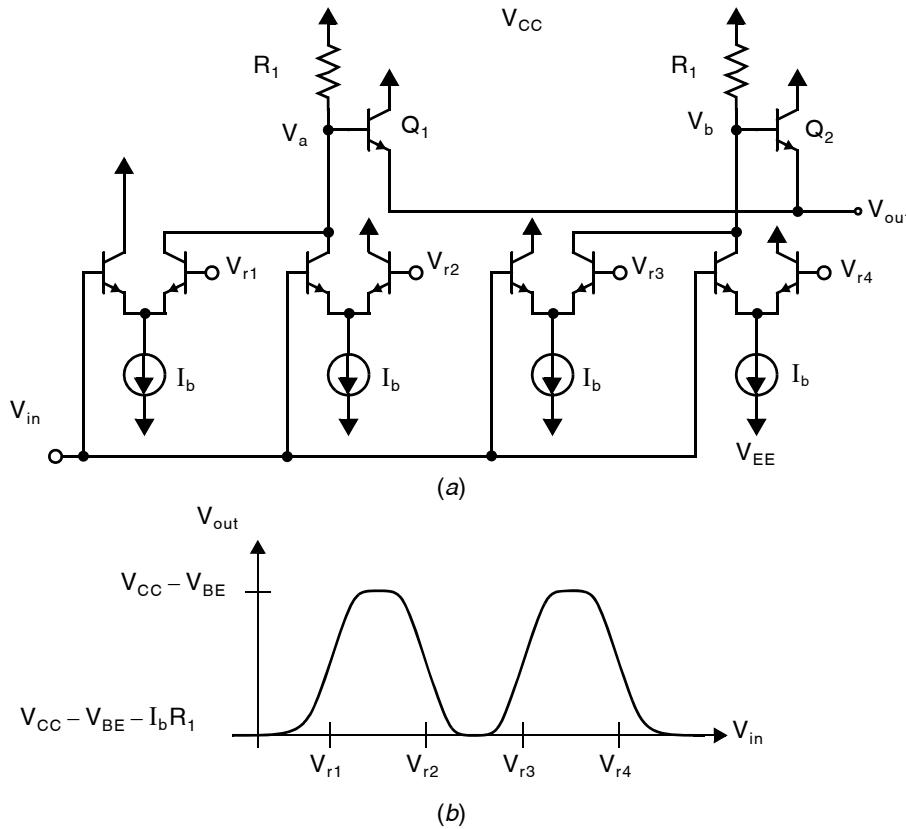


Fig. 17.36 A folding block with a folding-rate of four. (*a*) A possible single-ended circuit realization; (*b*) input–output response.

Some points worth mentioning here are that for full-scale input signals, the output signal from a folding block is at a much higher frequency than the input signal. In fact, the frequency of the folding block's output signal is equal to the multiplication of the frequency of the input signal times the folding rate. This multiplying effect limits the practical folding rate used in high-speed converters. Also, it should be mentioned that the circuit shown is a single-ended version, and differential circuits are almost always used in practical implementations.

Another point to note here is that while the folding approach reduces the number of latch comparators, a large input capacitance similar to that for a flash converter is also present with the folding circuit shown. In fact, flash converters have similar input stages of differential pairs of transistors for each comparator, but they are, of course, not cross coupled. Since the number of differential pairs in each folding block equals the folding rate, and the input signal goes to one side of each differential pair, it can be shown that the number of transistors driven by the input signal equals 2^N—the same number as for a flash converter. To reduce this large input capacitance, folding converters also make use of an interpolating architecture. With an interpolate-by-two technique applied to the 4-bit example, the resulting architecture would be that shown in Fig. 17.37. Note that a new inverted signal $\overline{V_4}$ is required to connect the top folding block to the bottom one. Although the creation of this inverted signal is needed in a single-ended version, no extra circuitry is required in a differential version since it can be accomplished by simply cross-coupling the differential output wires.

Key Point: *Folding is often combined with interpolation so that both the number of latches and the number of input amplifier differential pairs is reduced.*

In [van Valburg, 1992], in order to realize a very-high-speed 8-bit converter, four folding blocks each with a folding rate of eight were used. Assuming V_{ref} = 1 V, then the difference between reference voltages of adjacent inputs of a folding block was 1/8 V. Each adjacent folding block was offset by 1/32 V, giving zero crossings at each 1/32 V. By interpolating between each adjacent folding block using four 8-tap resistor strings, zero crossings every 1/256 V were obtained (i.e., an 8-bit converter with 32 latch comparators connected to the interpolating resistor strings). The MSBs were realized by taking appropriate outputs from selected differential pairs and summing them separately to realize additional folding amplifiers with reduced folding rates. The converter also included circuitry to prevent bubble errors, as shown in Fig. 17.27. Other examples of folding converters are given in [van de Grift, 1987, and Colleran, 1993]. In addition, folding converters have been used in the internal operation of a two-step converter [Vorenkamp, 1992].

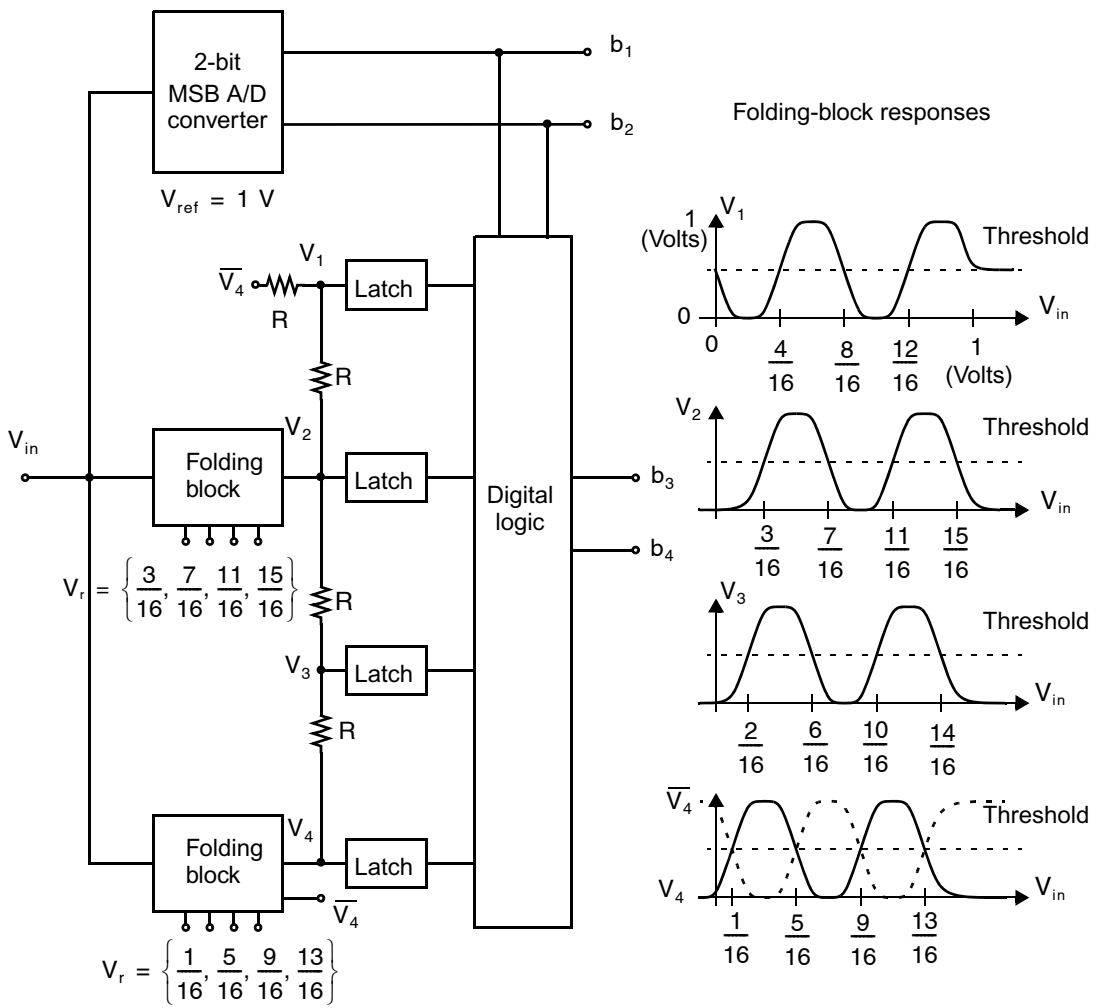


Fig. 17.37 A 4-bit folding A/D converter with a folding rate of four and an interpolate-by-two. (The MSB converter would usually be realized by combining some folding-block signals.)

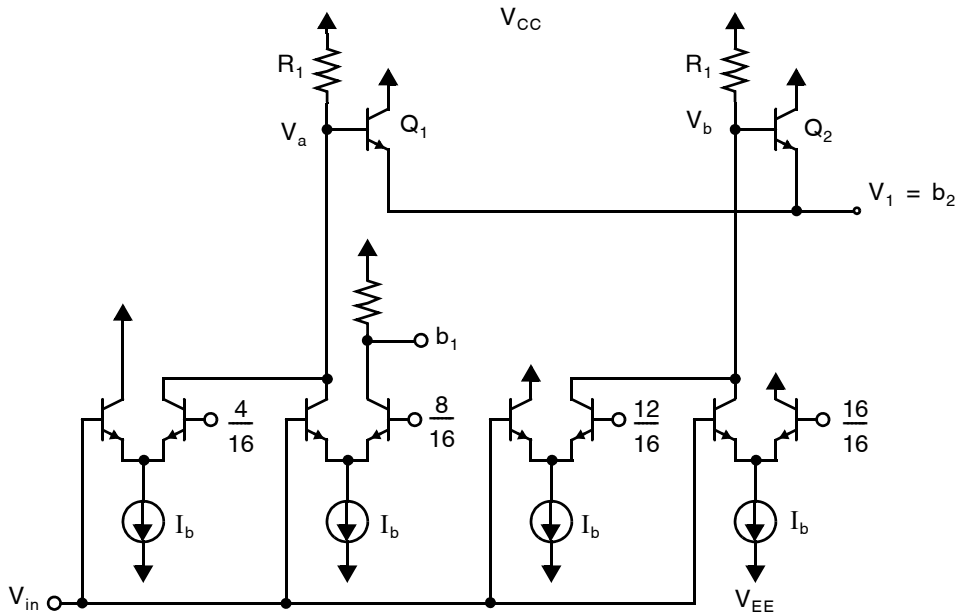


Fig. 17.38 Using the V_1 folding block to also determine the top two MSBs.

EXAMPLE 17.11

In Fig. 17.35, show how the two MSBs can be derived using internal signals from the folding blocks.

Solution

The MSB, b_1 is determined by the input signal being above or below $V_{ref}/2$. Looking at Fig. 17.35, we see that the input is compared to the appropriate signal, $(8/16)V_{ref}$ in the top folding block, producing V_1. As a result, the MSB is easily obtained by using the collector current of the transistor connected to V_{r2} in the top folding block. Similarly, the top folding block uses references $4/16$ and $12/16$, which are needed to determine the second bit, b_2. In fact, the signal V_1 can be used directly as the second bit, b_2, as seen in Fig. 17.35. Thus, the top two bits, b_1 and b_2, can be determined using the top folding block, as shown in Fig. 17.38.

17.9 TIME-INTERLEAVED A/D CONVERTERS

Very-high-speed A/D conversions can be realized by operating many A/Ds in parallel [Black, 1980]. The system architecture for a four-channel A/D is shown in Fig. 17.39. Here, ϕ_0 is a clock at four times the rate of ϕ_1 to ϕ_4. Additionally, ϕ_1 to ϕ_4 are delayed with respect to each other by the period of ϕ_0, such that each converter will get successive samples of the input signal, V_{in}, sampled at the rate of ϕ_0. In this way, the four A/D converters operate at one-quarter the rate of the input sampling frequency.

Key Point: *Time-interleaved A/D converters permit high sampling frequency using several lower-rate A/D converters and several phase-shifted clocks.*

With this approach, the input S/H making use of ϕ_0 is critical, while the remaining four S/H converters can have considerable jitter since the signal is already sampled at that point. Thus, sometimes the input S/H is realized in a different technology, such as GaAs, while the remaining S/H circuits could be realized in silicon. An example of a 1 GHz 6-bit A/D converter using time interleaving and GaAs S/H circuits is described in [Poulton, 1987], where four bipolar converters operating at 250 MHz were used.

Key Point: *A major limitation on the performance of time-interleaved converters is mismatch through the parallel signal paths in either their dc offset, gain, or sampling time. If they can be accurately quantified, such errors can be cancelled digitally.*

It is also essential that the different S/H and A/D converter signal paths are extremely well matched, as mismatches will produce tones. For example, consider a m-way time-interleaved converter where one converter has a dc offset of, say, 100 mV. Such a system will produce every m-th digital word different from the other $(m-1)$ and hence tones at f_s/m and its harmonics. These tones are independent of the input signal frequency or amplitude, and in fact will be present even if the input is zero. Another source of error arises when the different S/H and A/D converter circuits present different signal gains. In this case, since the gain is changing periodically, the input signal is effectively being multiplied by a periodic signal. Hence, with a sinusoidal input at f_{in} the output will contain harmonics at $kf_s/m \pm f_{in}$ for integers k. Because these are intermodulation products, their frequency and amplitude depends upon the input signal frequency and amplitude. Mismatch between the bandwidth of the parallel signal paths results in frequency-dependent gain mismatch and all of the attendant error tones. Such nonideal behavior can be disastrous for many applications since the tones may reside well within the frequency of interest. A sketch of the output spectrum in the presence of offset and gain mismatches is

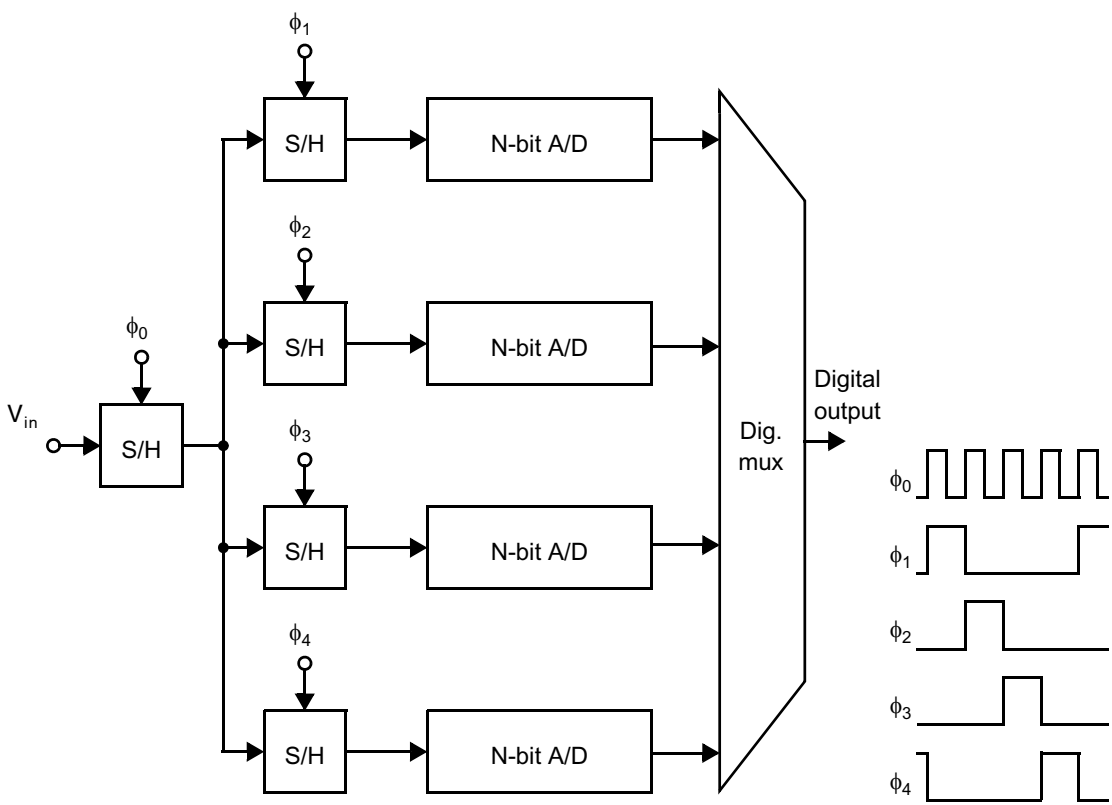


Fig. 17.39 A four-channel time-interleaved A/D converter with clock waveforms shown.

presented in Fig. 17.40 with the sources of the various tones identified. Fortunately, if the mismatches between signal paths can be accurately identified they can be cancelled digitally as in Fig. 17.41. Even when cancelled, offset and gain mismatches increase the dynamic range that the constituent A/D converters must be able to handle without nonlinearity.

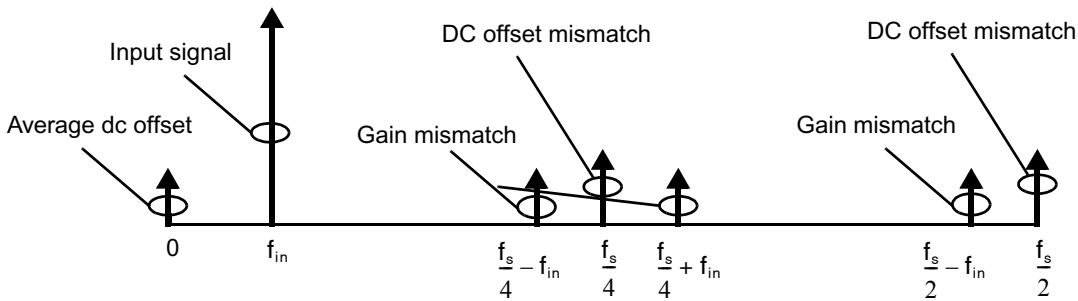


Fig. 17.40 Spectrum at the output of a 4-way time-interleaved A/D converter with offset and gain mismatches up to the converter's Nyquist frequency, $f_s/2$.

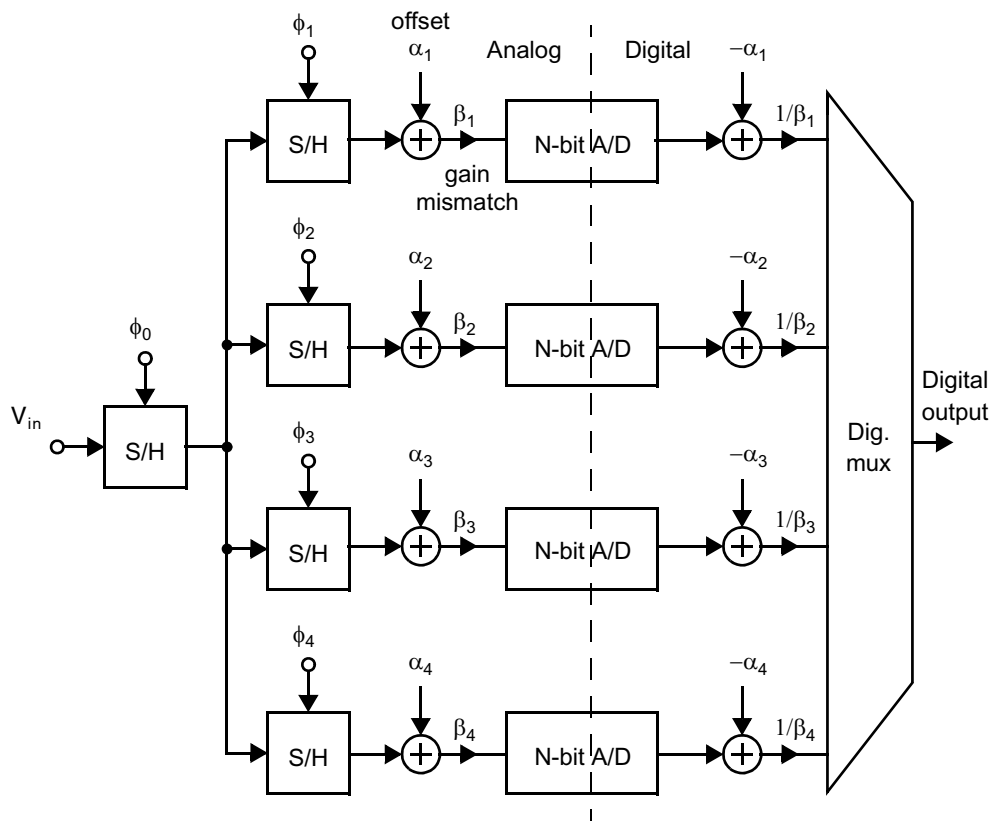


Fig. 17.41 A four-channel time-interleaved A/D converter with digital correction for gain and offset mismatch.

17.10 KEY POINTS

- Dual-slope integrating conversion proceeds first by integrating the input for a fixed time period, then applying a known input to the integrator and measuring the time required for the integrator output to return to zero. [p. 648]
- Dual-slope conversion does not require accuracy in defining the integration time constant. Offset errors can be calibrated by performing an additional conversion using a known dc input quantity. [p. 648]
- Integrating converters inherently filter the input with a sin x/x frequency response. [p. 649]
- Successive-approximation A/D converters are very versatile, capable of moderately high speed or accuracy with relatively low power. They are relatively simple circuits, in the simplest cases requiring only a single comparator, a bank of capacitors and switches, and a small digital logic circuit. Their biggest drawback is that they operate iteratively and therefore require many clock cycles to perform a single conversion. [p. 652]
- The accuracy of successive-approximation converters is often determined by that of the D/A converter. In the case of charge-redistribution converters, this role is played by a bank of comparators whose matching is important. [p. 653]
- Algorithmic converters are similar to successive-approximation converters, except that instead of an accurate D/A converter or capacitor bank, these converters require an accurate gain of two in order to perform the iterative search for a digital code that represents the analog input. [p. 662]
- Pipelined A/D converters are similar to algorithmic converters, but employ multiple circuits working on successive input samples simultaneously. Hence, they can complete a conversion on every clock cycle, providing higher throughput than algorithmic converters, but with the same latency. [p. 665]
- Introducing redundancy between successive stages in a pipelined converter permits simple digital correction of any offset in the comparators of each stage. [p. 670]
- Pipelined A/D converter performance is limited by nonlinearities and noise in the MDAC, and inaccuracy in the interstage gain, none of which are cancelled by simple digital correction. [p. 670]
- Switching part of the sampling capacitor into the feedback network provides the interstage gain with a higher feedback factor, and hence higher bandwidth, than a conventional switched-capacitor gain circuit. [p. 672]
- Multiple bits can be resolved in each stage, with digital correction, in order to reduce the number of opamps in a pipelined converter and decrease the number of clock cycles required for each conversion. [p. 673]
- Flash converters are generally capable of the fastest conversion speeds of any A/D architecture. However, their power consumption grows dramatically with the required resolution so they are usually reserved for converters where high speed and low or modest resolution is sought. [p. 673]
- Two-step converters may be considered a special case of pipelined converters where only 2 pipelined stages are needed. As with pipelined converters, digital error correction greatly relaxes the requirements on the first stage. [p. 679]
- The large capacitive input presented by flash A/D converters can be reduced by using fewer input amplifiers and interpolating between them to create the missing intermediate values. [p. 680]
- A folding architecture requires fewer latches than a flash converter. It does so similar to a two-step converter, but uses analog preprocessing instead of an accurate D/A converter and gain amplifier. [p. 683]
- Folding is often combined with interpolation so that both the number of latches and the number of input amplifier differential pairs is reduced. [p. 686]
- Time-interleaved A/D converters permit high sampling frequency using several lower-rate A/D converters and several phase-shifted clocks. [p. 687]
- A major limitation on the performance of time-interleaved converters is mismatch through the parallel signal paths in either their dc offset, gain, or sampling time. If they can be accurately quantified, such errors can be cancelled digitally. [p. 688]

17.11 REFERENCES

Analog Devices. *Analog-Digital Conversion Handbook*. Prentice Hall, Englewood Cliffs, New Jersey, 1986.

W. C. Black, Jr., and D. A. Hodges. "Time-Interleaved Converter Arrays," *IEEE J. of Solid-S tate Cir cuits*, Vol. 15, pp. 1022–1029, December 1980.

W. T. Colleran, T. H. Phan, and A. A. Abidi. "A 10-b, 100-Ms/s Pipelined A/D Converter," *IEEE Int. Solid-State Circuits Conf.,* pp. 68–69. San Francisco, February 1993.

A. G. F. Dingwall. "Monolithic Expandable 6-bit 20-MHz CMOS/SOS A/D Converter," *IEEE J. of Solid-S tate Cir cuits*, Vol. 14, pp. 926–931, December 1979.

B. Fotouhi and D. A. Hodges. "High-Resolution A/D Conversion in MOS/LSI," *IEEE J. of So lid-State Circuits*, Vol. 14, pp. 920–925, December 1979.

Y. Gendai, Y. Komatsu, S. Hirase, and M. Kawata. "An 8-b 500-MHz ADC," *IEEE Int. Solid-State Cir cuits Conf.,* pp. 172–173, San Francisco, February 1991.

F. Goodenough. "Interpolators Put 10-bit 75-MHz A/D Converters on 8-bit Digital Process," *Electronic D esign*, pp. 29–30, December 14, 1989.

M. Ito et al. "A 10b 20Ms/s 3V-Supply CMOS A/D Converter for Integration Into System VLSIs," *IEEE Int. Solid-S tate Circuits Conf.*, pp. 48–49, San Francisco, February 1994.

K. Kusumoto et al. "A 10b 20MHz 30mW Pipelined Interpolating CMOS ADC," *IEEE Int. Solid-State Circuits Conf.,* pp. 62–63, San Francisco, February 1993.

H. Lee, D. A. Hodges, and P. R. Gray. "A Self-Calibrating 15-bit CMOS A/D Converter," *IEEE J. of Solid-S tate Cir cuits*, Vol. 19, pp. 813–819, December 1984.

S. H. Lewis, H. S. Fetterman, G. F. Gross, R. Ramachandran, T. R. Viswanathan, "A 10-b 20-Msample/s Analog-to-Digital Converter," *IEEE J. of Solid-State Circuits,* Vol. 27, pp. 351–358, March 1992.

P. W. Li, M. J. Chin, P. R. Gray, and R. Castello. "A Ratio-Independent Algorithmic Analog-To-Digital Conversion Technique," *IEEE J. of Solid-State Circuits,* Vol. 19, pp. 828–836, December 1984.

K. W. Martin. "A High-Speed, High-Accuracy Pipelined A/D Converter," *IEEE 15 th A silomar C onf. on Cir cuits, Systems, an d Computers*, November 1981.

R. H. McCharles, V. A. Saletore, W. C. Black, Jr., and D. A. Hodges. "An Algorithmic Analog-To-Digital Converter," *I EEE In t. Solid-State Circuits Conf.*, Philadelphia, February 1977.

J. L. McCreary et al. "All-MOS Charge Redistribution A/D Conversion Technique — Part I," *IEEE J. of Solid-State Circuits*, Vol. 10, pp. 371–379, December 1975.

R. Petschacher et al. "A 10-b 75-MSPS Subranging A/D Converter with Integrated Sample and Hold," *IEEE J. of Solid-State Cir cuits*, Vol. 25, pp. 1339–1346, December 1990.

K. Poulton, J. J. Corcoran, and T. Hornak. "A 1-GHz 6-bit ADC System," *IEEE J. of Solid-State C ircuits,* Vol. 22, pp. 962–970, December 1987.

A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 3rd ed. Saunders College Publishing/HRW, New Jersey, 1991.

M. Steyaert, R. Roovers, and J. Craninckx. "A 100-MHz 8-bit CMOS Interpolating A/D Converter," *IEEE Custom Integrated Cir cuits Conf.*, pp. 28.1.1–28.1.4, San Diego, May 1993.

S. Sutarja, "High-resolution pipelined analog-to-digital conversion," Ph.D. Thesis, Univ. California Berkeley, pp. 40–42, May 1988.

K-S. Tan et al. "Error Correction Techniques for High-Performance Differential A/D Converters," *IEEE J. of Solid-S tate Cir cuits,* Vol. 25, pp. 1318–1326, December 1990.

R. E. J. van de Grift, I. W. J. M. Rutten, and M. van der Veen. "An 8-bit Video ADC Incorporating Folding and Interpolation Techniques," *IEEE J. of Solid-State Circuits*, Vol. 22, pp. 944–953, December 1987.

R. J. van de Plassche and P. Baltus. "An 8-bit 100-MHz Full-Nyquist Analog-to-Digital Converter," *IEEE J. of Solid-S tate C ir cuits*, Vol. 23, pp. 1334–1344, December 1988.

J. van Valburg and R. J. van de Plassche. "An 8-b 650-MHz Folding ADC," *IEEE J. of Solid-S tate Cir cuits*, Vol. 27, pp. 1662–1666, December 1992.

B. Verbruggen, J. Craninckx, M. Kuijk, P. Wambacq, G. Van der Plas, "A 2.2 mW 1.75 GS/s 5 Bit Folding Flash ADC in 90 nm Digital CMOS," vol. 44, pp. 874–882, March 2009.

P. Vorenkamp, and J. P. M. Verdaasdonk. "A 10b 50MS/s Pipelined ADC," *IEEE Int. Solid-State Cir cuits Conf.*, pp. 32–33, San Francisco, February 1992.

17.12 PROBLEMS

17.12.1 Section 17.1: Integrating Converters

17.1 What is the worst-case conversion time for an 18-bit dual-slope integrating A/D converter when the clock rate is 5 MHz?

17.2 Consider an 18-bit integrating A/D converter, as shown in Fig. 17.1, where V_{ref} equals 10 volts, $C_1 = 100$ pF, and a clock frequency of 1 MHz is used. What value of R_1 should be chosen such that the opamp output never exceeds 10 volts when $0 V < V_{in} < 10$ V?

17.3 Derive an expression equivalent to (17.8) for Fig. 17.1 when the opamp has an input-offset voltage of V_{off1} and the comparator has an input-offset voltage of V_{off2}.

17.4 What is the offset error (in LSBs) of the converter described in Problem 17.2 when the integrating opamp has an input-offset voltage of 20 mV?

17.5 What input-signal frequencies are completely attenuated by a dual-slope 16-bit integrating A/D having a clock frequency of 1 MHz? For this same converter, what is the attenuation of an input signal at 60 Hz?

17.6 Repeat Problem 17.5 with the converter's clock frequency being equal to 100 kHz.

17.12.2 Section 17.2: Successive-Approximation Converters

17.7 Consider a 4-bit unipolar DAC-based successive-approximation A/D converter, as shown in Fig. 17.5. Find the sequence of the D/A converter's output levels for an input of 3.333 volts when $V_{ref} = 8$ V. What is the final digital output?

17.8 Consider a 4-bit unipolar charge-redistribution A/D converter, as shown in Fig. 17.7, where $V_{ref} = 8$ V. Find the sequence of the voltage level for V_x if the input signal is 3.333 volts.

17.9 Repeat Problem 17.8, assuming that a parasitic capacitance of 8 C is connected between the node at V_x and ground. Is the final digital result affected?

17.10 Show a method for modifying the unipolar charge-redistribution A/D converter shown in Fig. 17.7 such that the threshold offsets of $0.5V_{LSB}$ are correctly realized.

17.11 How is the digital code obtained from the signed charge redistribution A/D converter shown in Fig. 17.8 converted to a 2's complement code?

17.12 What value of error voltage V_{x1} would be measured in Fig. 17.12(*a*) if the total array capacitance equals 64 pF, the MSB (i.e., b_1) capacitor equals 31.5 pF, and there is a parasitic capacitance to ground of 10 pF on node V_x? For the same circuit, show that the error voltage, V_{e1}, that occurs when using the MSB capacitor during a normal conversion equals $0.5V_{x1}$.

17.13 For the same circuit as described in Problem 17.12, what value of error voltage V_{x2} would be measured in Fig. 17.12(*b*) if the MSB-1 (i.e., b_2) capacitor equals 16.4 pF? Show that the error voltage, V_{e2}, that occurs when using the MSB-1 capacitor during a normal conversion equals $0.5(V_{x2} - V_{e1})$.

17.14 Assuming switch resistances are all about 1 kΩ, estimate the settling time needed for a 12-bit charge-redistribution A/D converter where the total array capacitance equals 128 pF.

17.12.3 Section 17.3: Algorithmic (or Cyclic) A/D Converter

17.15 What is the latency in a 12-bit algorithmic A/D converter? That is, how many clock cycles are required in order to perform a complete conversion?

17.12.4 Section 17.4: Pipelined A/D Converters

17.16 In a 9-bit 1-bit-per-stage pipelined A/D converter with a full-scale input range of $-1\ \text{V} < V_{in} < 1\ \text{V}$, what is the maximum tolerable offset in the first stage comparator while ensuring the output has less than 1/2 LSB overall error? If the converter is changed to a 1.5-bit-per-stage architecture with error correction, how much offset can be tolerated in the first stage comparators? You may assume that the 2nd and all subsequent stages are ideal.

17.17 In a 11-bit 1.5-bit-per-stage pipelined converter, what is the required accuracy of the reference voltages $\pm V_{ref}/2$ in the first stage of the pipeline? What about the fourth stage in the pipeline?

17.18 Draw a block diagram for a 4-bit, 1.5-bit-per-stage pipelined converter with error correction. With $V_{ref} = 1\ \text{V}$ and $V_{in} = 75\ \text{mV}$, find the residue voltages V_2 and V_3, the digital codes at the outputs of each stage, and the final digital output after decoding assuming all blocks are ideal. Repeat assuming the upper comparator of stage 1 makes an error.

17.19 A 10-bit pipelined converter is to be designed using only 5 stages. What should the interstage gain be?

17.20 Modify the circuit of Fig. 17.22 to have the input-output relationship shown in Fig. P17.20 as would be required for a 2.5-bit-per-stage pipelined converter.

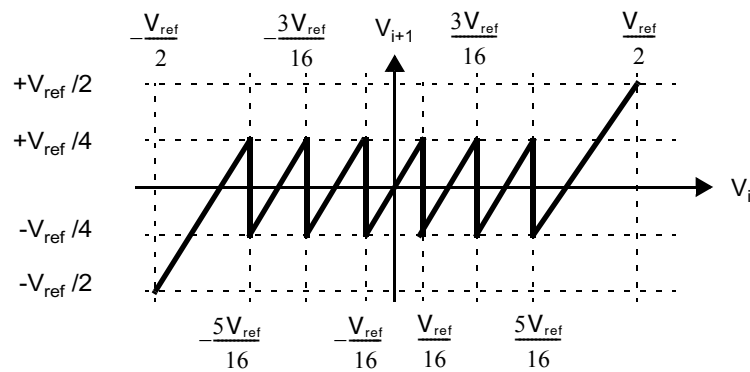


Fig. P17.20

17.12.5 Section 17.5: Flash Converters

17.21 Consider the clocked comparator shown in Fig. 17.25 in the 0.18-μm technology of Table 1.5 and assuming

$$2\left(\frac{W}{L}\right)_n = \left(\frac{W}{L}\right)_p = 2$$

$$V_{DD} = 1.8\ \text{V}$$

If the MOSFETS in the inverter are ideal except for each having a 10-kΩ output impedance when active, find the minimum differential input that will cause a 0.2-V change (use a linear analysis). Assuming this minimum differential input corresponds to $(1/2)$ LSB, what is the maximum number of bits that this comparator could determine in a flash A/D converter?

17.22 Assume each comparator in the 3-bit flash converter of Fig. 17.24 draws an input dc bias current of 0.1 μA from the reference voltage supply V_{ref}. Find the resistor value R required to ensure less than 1% error in the reference ladder voltages.

17.12.6 Section 17.6: Two-Step A/D Converters

17.23 Draw a block diagram similar to that for Fig. 17.30 for a 10-bit two-step A/D converter where the first stage determines 4 bits. Indicate the accuracy needed in all the blocks.

17.24 Consider the two-step converter with digital error correction in Fig. 17.30. The input signal is $V_{in} = 20$ mV and the converter's full-scale range is $0 < V_{in} < 1$ V. The output of the MSB A/D converter is 0001 and all other components in the converter are ideal. Find the signals at V_1, V_q, and the 5-bit output of the LSB A/D. Show that the error correction circuitry produces the correct output code.

17.12.7 Section 17.7: Interpolating A/D Converters

17.25 In Section 17.7, it was stated that the interpolated signals will have the correct threshold crossings if V_1 and V_2 in Fig. 17.32 are linear in the region $0.25 < V_{in} < 0.5$. In fact, this region is larger than what is actually necessary. What is the minimum region of V_{in} that V_1 and V_2 should be linear over?

17.26 Many very-high-speed A/D converters do not use a sample and hold since it would limit their speed, but instead allow the input signal to be applied to the array of comparators, which are all clocked "simultaneously." Qualitatively, explain why a clocked flash or interpolating A/D converter might operate faster than a sample-and-hold circuit.

17.27 Show that the circuit shown in Fig. 17.33 results in all time constants being equal when each of the latches has the same input capacitance and the amplifiers have zero output impedance.

17.12.8 Section 17.8: Folding A/D Converters

17.28 In an N-bit folding A/D converter (with no interpolation), what is the product of the folding rate times the number of folding blocks?

17.29 Find an expression for the number of latches in an N-bit folding A/D converter where the folding rate is $FR = 2^l$.

17.30 Assuming the input capacitance of a differential stage is the same in either a flash or folding/interpolating A/D converter, what reduction of input capacitance over a flash converter would be achieved with an 8-bit folding/interpolating A/D converter having four folding blocks, each with a folding rate of eight? If a straight interpolating A/D converter is to have the same reduction in input capacitance, how many resistors are required between "input comparators"?

17.12.9 Section 17.9: Time-Interleaved A/D Converters

17.31 A time-interleaved converter with $m = 2$ signal paths is clocked at 500 MHz and driven with a sinusoidal input at 181 MHz. The output spectrum is shown in Fig. P17.31 up to the Nyquist frequency, 250 MHz. It includes

tones due to dc offset and gain mismatch in the time-interleaved architecture, as well as the 3rd harmonic of the input signal due to nonlinearity. Identify the source of each tone in the spectrum.

Fig. P17.31

17.32 Sketch the spectrum at the output of a 3-way time-interleaved A/D converter clocked at $f_s = 3$ GHz with $f_{in} = 800$ MHz. Assume both dc offset and gain mismatches are present, as well as 3rd-order distortion of the input.

CHAPTER

18 Oversampling Converters

Key Point: *Oversampling converters relax the requirements placed on the analog circuitry at the expense of more complicated digital circuitry.*

Oversampling A/D and D/A converters are popular for high-resolution medium-to-low-speed applications such as high-quality digital audio and baseband signal processing in some wireless systems. A major reason for their popularity is that oversampling converters relax the requirements placed on the analog circuitry at the expense of more complicated digital circuitry. This tradeoff became desirable with the advent of deep submicron CMOS technologies as complicated high-speed digital circuitry became more easily realized in less area, but the realization of high-resolution analog circuitry was complicated by the low power-supply voltages and poor transistor output impedance caused by short-channel effects. With oversampling data converters, the analog components have reduced requirements on matching tolerances and amplifier gains. Oversampling converters also simplify the requirements placed on the analog anti-aliasing filters for A/D converters and smoothing filters for D/A converters. For example, usually only a first- or second-order anti-aliasing filter is required for A/D converters, which can often be realized very inexpensively. Furthermore, a sample-and-hold is usually not required at the input of an oversampling A/D converter.

In this chapter, the basics of oversampling converters are discussed first. We shall see that extra bits of resolution can be extracted from converters that sample much faster than the Nyquist rate. Furthermore, this extra resolution can be obtained with lower oversampling rates by spectrally shaping the quantization noise through the use of feedback. The use of shaped quantization noise applied to oversampling signals is commonly referred to as delta-sigma ($\Delta\Sigma$) modulation.[1] Simple first- and second-order $\Delta\Sigma$ modulators are discussed, followed by a discussion of typical system architectures for $\Delta\Sigma$ data converters. Next, two popular approaches for realizing decimation filters are described. Descriptions of some modern approaches are then described along with some practical considerations. The chapter concludes with an example design of a third-order $\Delta\Sigma$ A/D converter.

18.1 OVERSAMPLING WITHOUT NOISE SHAPING

In this section, the advantage of sampling at higher than the Nyquist rate is discussed. Here, we shall see that extra dynamic range can be obtained by spreading the quantization noise power over a larger frequency range. However, we shall see that the increase in dynamic range is only 3 dB for every doubling of the sample rate. To obtain much higher dynamic-range improvements as the sampling rate is increased, noise shaping through the use of feedback can be used and is discussed in the next section.

1. Delta-sigma modulation is also sometimes referred to as sigma-delta modulation.

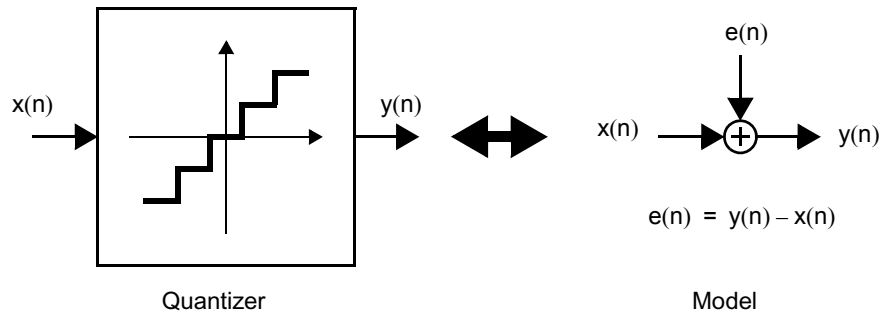


Fig. 18.1 Quantizer and its linear model.

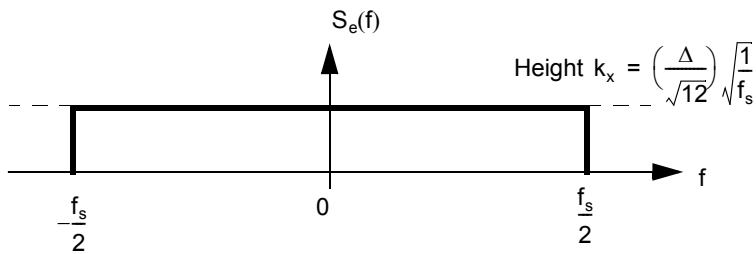


Fig. 18.2 Assumed spectral density of quantization noise.

18.1.1 Quantization Noise Modelling

We begin by modelling a quantizer as adding quantization error $e(n)$, as shown in Fig. 18.1. The output signal, $y(n)$, is equal to the closest quantized value of $x(n)$. The *quantization error* is the difference between the input and output values. This model is exact if one recognizes that the quantization error is not an independent signal but may be strongly related to the input signal, $x(n)$. This linear model becomes approximate when assumptions are made about the statistical properties of $e(n)$, such as $e(n)$ being an independent white-noise signal. However, even though approximate, it has been found that this model leads to a much simpler understanding of $\Delta\Sigma$ and with some exceptions is usually reasonably accurate.

18.1.2 White Noise Assumption

If $x(n)$ is very active, $e(n)$ can be approximated as an independent random number uniformly distributed between $\pm\Delta/2$, where Δ equals the difference between two adjacent quantization levels. Thus, the quantization noise power equals $\Delta^2/12$ (from Section 15.3) and *is independent of the sampling frequency*, f_s. Also, the spectral density of $e(n)$, $S_e(f)$, is white (i.e., a constant over frequency) and all its power is within $\pm f_s/2$ (a two-sided definition of power).

Assuming white quantization noise, the spectral density of the quantization noise, $S_e(f)$ appears as shown in Fig. 18.2.

Key Point: *Quantization with a step size "LSB" can be modeled as an additive quantization noise uniformly distributed between –LSB/2 and +LSB/2 with a white power spectrum and total power of* $LSB^2/12$.

The spectral density height is calculated by noting that the total noise power is $\Delta^2/12$ and, with a two-sided definition of power, equals the area under $S_e(f)$ within $\pm f_s/2$, or mathematically,

$$\int_{-f_s/2}^{f_s/2} S_e^2(f)\,df = \int_{-f_s/2}^{f_s/2} k_x^2\,df = k_x^2 f_s = \frac{\Delta^2}{12} \tag{18.1}$$

Solving this relation gives

$$k_x = \left(\frac{\Delta}{\sqrt{12}}\right)\sqrt{\frac{1}{f_s}} \tag{18.2}$$

EXAMPLE 18.1

Find the output and quantization errors for two different quantizers, as shown in Fig. 18.3, when the input values are

$$x(n) = \{0.01,\ 0.31,\ -0.11,\ 0.80,\ 0.52,\ -0.70\} \tag{18.3}$$

Also find the expected power and the power density height, $S_e^2(f)$, of the quantization noise when the sampling frequency is normalized to 2π rad/sample.

Solution

The output and quantization noise values for this example are given in Table 18.1. Note that although the output signals, $y(n)$, are well-defined values, the quantization error values can be approximated as uniformly distributed random numbers since the input signal is quite active.

Recalling that the quantization noise power is given by

$$P_e = \frac{\Delta^2}{12} \tag{18.4}$$

then the expected noise powers of the two quantizers are

$$P_I = \frac{0.5^2}{12} = 0.0208\ \text{W} \tag{18.5}$$

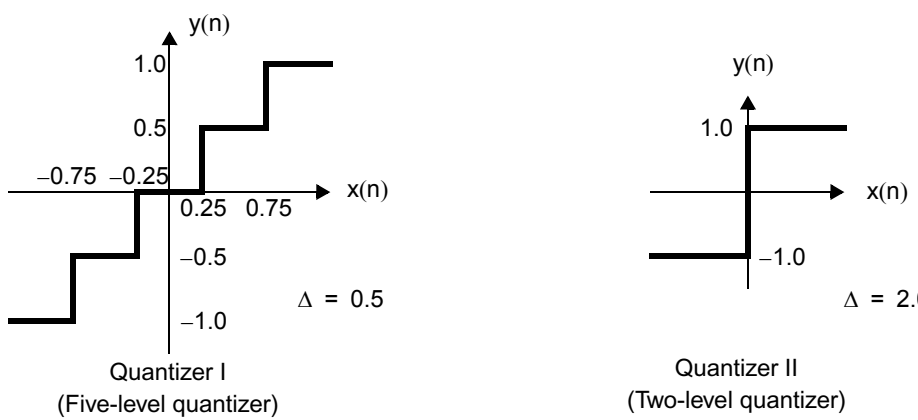


Fig. 18.3 Two example quantizers.

Table 18.1 Example signal values and quantization noise for two quantizers.

	Quantizer I (Five Levels)		Quantizer II (Two Levels)	
x(n)	y(n)	e(n)	y(n)	e(n)
0.01	0.0	–0.01	1	0.99
0.31	0.5	0.19	1	0.69
–0.11	0.0	0.11	–1	–0.89
0.80	1.0	0.2	1	0.20
0.52	0.5	–0.02	1	0.48
–0.70	–0.5	0.2	–1	–0.30

$$P_{II} = \frac{2^2}{12} = 0.333 \text{ W} \tag{18.6}$$

where we note that these values are not affected by normalizing the sampling frequency. Also note that the two power estimates for P_I and P_{II} correspond to rms power levels of 0.144 and 0.577, respectively.

For the power density, we note that the power is spread evenly between $\pm f_s/2$, resulting in a density of

$$S_e^2(f) = \frac{P_e}{f_s} \tag{18.7}$$

which results in

$$S_{eI}^2(f) = \left(\frac{0.0208}{2\pi}\right) = 0.00331 \frac{\text{W}}{\text{rad/sample}} \tag{18.8}$$

$$S_{eII}^2(f) = \left(\frac{0.333}{2\pi}\right) = 0.053 \frac{\text{W}}{\text{rad/sample}} \tag{18.9}$$

18.1.3 Oversampling Advantage

Oversampling occurs when the signals of interest are bandlimited to f_0 yet the sample rate is at f_s, where $f_s > 2f_0$ ($2f_0$ being the Nyquist rate or, equivalently, the minimum sampling rate for signals bandlimited to f_0). We define the oversampling ratio, OSR, as

$$\text{OSR} \equiv \frac{f_s}{2f_0} \tag{18.10}$$

After quantization, since the signals of interest are all below f_0, $y_1(n)$ is filtered by $H(f)$ to create the signal $y_2(n)$, as shown in Fig. 18.4. This filter eliminates quantization noise (together with any other signals) greater than f_0.

Assuming the input signal is a sinusoidal wave, its maximum peak value without clipping is $2^N(\Delta/2)$. For this maximum sinusoidal wave, the signal power, P_s, has a power equal to

$$P_s = \left(\frac{\Delta 2^N}{2\sqrt{2}}\right)^2 = \frac{\Delta^2 2^{2N}}{8} \tag{18.11}$$

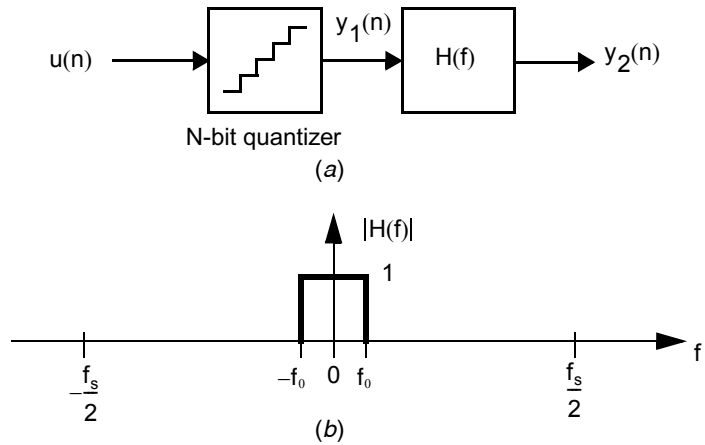


Fig. 18.4 (*a*) A possible oversampling system without noise shaping. (*b*) The brick-wall response of the filter to remove much of the quantization noise.

The power of the input signal within $y_2(n)$ remains the same as before since we assumed the signal's frequency content is below f_0. However, the quantization noise power is reduced to

$$P_e = \int_{-f_s/2}^{f_s/2} S_e^2(f)|H(f)|^2\,df = \int_{-f_0}^{f_0} k_x^2\,df = \frac{2f_0}{f_s}\frac{\Delta^2}{12} = \frac{\Delta^2}{12}\left(\frac{1}{OSR}\right) \tag{18.12}$$

Therefore, doubling OSR (i.e., sampling at twice the rate) decreases the quantization noise power by one-half or, equivalently, 3 dB (or, equivalently, 0.5 bits).

We can also calculate the maximum SQNR (in dB) to be the ratio of the maximum sinusoidal power to the quantization noise power in the signal $y_2(n)$. Mathematically, we have through the use of (18.11) and (18.12)

$$SQNR_{max} = 10\log\left(\frac{P_s}{P_e}\right) = 10\log\left(\frac{3}{2}2^{2N}\right) + 10\log(OSR) \tag{18.13}$$

which is also equal to

$$SQNR_{max} = 6.02N + 1.76 + 10\log(OSR) \tag{18.14}$$

Key Point: *Oversampling a signal at OSR times the Nyquist rate, without spectrally shaping the quantization noise, results in a SNR enhancement of 10 log(OSR) dB.*

The first term is the SQNR due to the N-bit quantizer while the OSR term is the SQNR enhancement obtained from oversampling. Here we see that straight oversampling gives a SQNR improvement of 3 dB/octave or, equivalently, 0.5 bits/octave. The reason for this SQNR improvement through the use of oversampling is that when quantized samples are averaged together, the signal portion adds linearly, whereas the noise portion adds as the square root of the sum of the squares. Note that this enhancement also takes place with respect to other types of noise, such as thermal noise in circuits. Hence, oversampling generally improves overall signal to noise ratio (SNR) by 10 log(OSR).

EXAMPLE 18.2

Consider a sampled dc signal, V_s, of value 1 V where the measured voltage, V_{meas}, is V_s plus a noise signal, V_{noise}. Assume V_{noise} is a random signal uniformly distributed between $\pm\sqrt{3}$. What is the SNR for V_{meas} when looking at individual values? If eight samples of V_{meas} are averaged together, roughly what is the new SNR? To illustrate the results, use eight typical samples for V_{meas} of {0.94, –0.52, –0.73, 2.15, 1.91, 1.33, –0.31, 2.33}.

Solution

Referencing signals to $1\ \Omega$, we calculate the power of V_s and V_{noise} to both be 1 watt. Thus the SNR for V_{meas} is 0 dB when looking at individual V_{meas} values. Note that it is difficult to see the signal value of 1 V in the example V_{meas} samples since the SNR is so poor.

If eight samples are averaged, we are realizing a modest low-pass filter, resulting in the oversampling ratio being approximately equal to 8 (this is a rough estimate since a brick-wall filter is not being used). Since each octave of oversampling results in a 3-dB SNR improvement, the averaged value should have a SNR of around 9 dB. Note that averaging the eight given V_{meas} samples results in 0.8875, which more closely represents the signal value of 1 V.

The reason oversampling improves the SNR here is that by summing eight measured values, the eight signal values add linearly to 8 V_{rms} (or 64 watts) while the eight noise values add to $\sqrt{8}\ V_{rms}$ (or 8 watts), since the noise values are assumed to be independent.

EXAMPLE 18.3

Given that a 1-bit A/D converter has a 6-dB SQNR, what sample rate is required using oversampling (no noise shaping) to obtain a 96-dB SQNR (i.e., 16 bits) if $f_0 = 25$ kHz? (Note that the input into the A/D converter has to be very active for the white-noise quantization model to be valid—a difficult arrangement when using a 1-bit quantizer with oversampling without noise shaping.)

Solution

Oversampling (without noise shaping) gives 3 dB/octave where 1 octave implies doubling the sampling rate. We require 90 dB divided by 3 dB/octave, or 30 octaves. Thus, the required sampling rate, f_s, is

$$f_s = 2^{30} \times 2f_0 \cong 54{,}000 \text{ GHz!}$$

This example shows why noise shaping is needed to improve the SQNR faster than 3 dB/octave, since 54,000 GHz is highly impractical.

18.1.4 The Advantage of 1-Bit D/A Converters

While oversampling improves the signal-to-noise ratio, it does not improve linearity. For example, if a 16-bit linear converter is desired while using a 12-bit converter with oversampling, the 12-bit converter must have an integral nonlinearity error less than $1/2^4$ LSB (here, LSB refers to that for a 12-bit converter). In other words, the component accuracy would have to match better than 16-bit accuracy (i.e., $100 \times (1/2^{16}) = 0.0015$ percent

Key Point: *One-bit D/A converters are inherently linear. Hence, when combined with a high oversampling ratio, they can provide a good signal to quantization noise and distortion ratio.*

accuracy). Thus, some sort of auto calibration or laser trimming must be used to obtain the required linearity. However, as we saw in Example 18.3, with a high enough sampling rate, the output from a 1-bit converter can be filtered to obtain the equivalent of a 16-bit converter. *The advantage of a 1-bit D /A is that it is inherently linear.*[2] This linearity is a result of a 1-bit D/A converter having only two output values and, since two points define a straight line, no trimming or calibration is required. This inherent linearity is one of the major motivations for making use of oversampling techniques with 1-bit converters. In fact, the reader may be aware that many audio converters use 1-bit converters for realizing 16- to 18-bit linear converters (with noise shaping). In fact, 20-bit linearity has been reported without the need for any trimming [Leopold, 1991]. Finally, it should be mentioned that there are other advantages when using oversampling techniques, such as a reduced requirement for analog anti-aliasing and smoothing filters.

18.2 OVERSAMPLING WITH NOISE SHAPING

In this section, the advantage of noise shaping the quantization noise through the use of feedback is discussed. Here, we shall see a much more dramatic improvement in dynamic range when the input signal is oversampled as compared to oversampling the input signal with no noise shaping. Although this section illustrates the basic principles with reference to A/D converters, much of it applies directly to ΔΣ D/A converters as well.

The system architecture of a ΔΣ oversampling A/D converter is shown in Fig. 18.5. The first stage is a continuous-time anti-aliasing filter and is required to band-limit the input signal to frequencies less than one-half the oversampling frequency, f_s. When the oversampling ratio is large, the anti-aliasing filter can often be quite simple, such as a simple RC low-pass filter. Following the anti-aliasing filter, the continuous-time signal, $x_c(t)$, is sampled by a sample and hold. This signal is then processed by the ΔΣ modulator, which converts the analog signal into a noise-shaped low-resolution digital signal. The third block in the system is a decimator. It converts the oversampled low-resolution digital signal into a high-resolution digital signal at a lower sampling rate usually equal to twice the frequency of the desired bandwidth of the input signal. The decimation filter can be conceptually thought of as a low-pass filter followed by a down sampler, although in many systems the decimation is performed in a number of stages. It should be mentioned that in many realizations where the ΔΣ modulator is realized using switched-capacitor circuitry, a separate sample-and-hold is not required, as the continuous-time signal

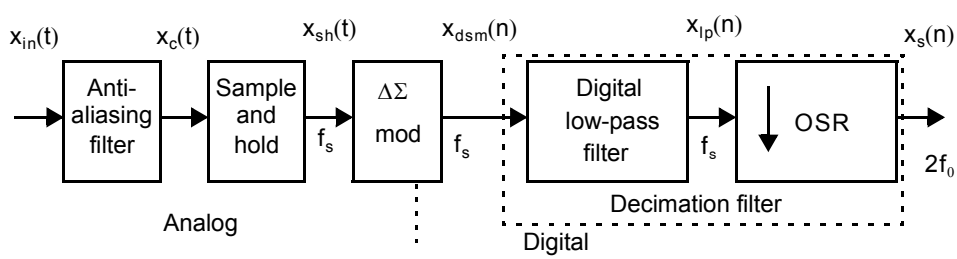


Fig. 18.5 Block diagram of an oversampling A/D converter.

2. This assumes that second-order errors due to imperfections such as signal-dependent power supply, or reference voltages, or memory in the D/A converter are not present.

is inherently sampled by the switches and input capacitors of the SC ΔΣ. In the next few sections the operation of the various building blocks will be described in greater detail.

18.2.1 Noise-Shaped Delta-Sigma Modulator

A general noise-shaped delta-sigma (ΔΣ) modulator and its linear model are shown in Fig. 18.6. This arrangement is known as an interpolative structure and is analogous to an amplifier realized using an opamp and feedback. In this analogy, the feedback reduces the effect of the noise of the output stage of the opamp in the closed-loop amplifier's output signal at low frequencies when the opamp gain is high. At high frequencies, when the opamp's gain is low, the noise is not reduced. Note that the quantizer here is shown for the general case where many output levels occur. While many oversampling converters make use of 1-bit quantizers (i.e., only two output levels) due to reasons already discussed, there is certainly no reason to restrict ourselves to such implementations. Multibit oversampling converters are discussed in detail in Section 18.8.

Treating the linear model shown in Fig. 18.6(*b*) as having two *independent* inputs (which is an approximation), we can derive a signal transfer function, $S_{TF}(z)$, and a noise transfer function, $N_{TF}(z)$.

$$S_{TF}(z) \equiv \frac{Y(z)}{U(z)} = \frac{H(z)}{1 + H(z)} \tag{18.15}$$

$$N_{TF}(z) \equiv \frac{Y(z)}{E(z)} = \frac{1}{1 + H(z)} \tag{18.16}$$

Note that the zeros of the noise transfer function, $N_{TF}(z)$, will be equal to the poles of $H(z)$. In other words, when $H(z)$ goes to infinity, we see by (18.16) that $N_{TF}(z)$ will go to zero. We can also write the output signal as the combination of the input signal and the noise signal, with each being filtered by the corresponding transfer function. In the frequency domain, we have

$$Y(z) = S_{TF}(z)U(z) + N_{TF}(z)E(z) \tag{18.17}$$

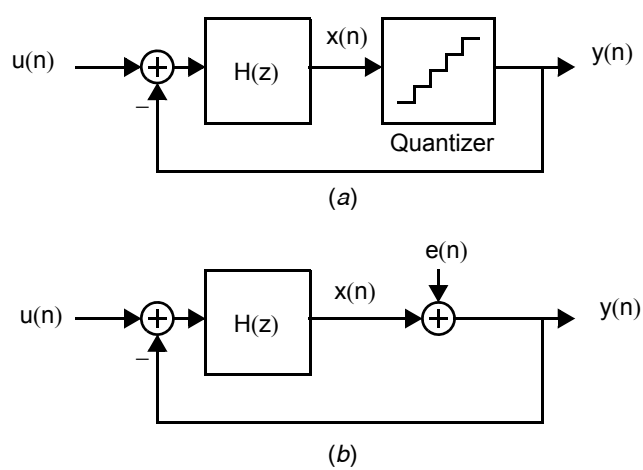


Fig. 18.6 A modulator and its linear model: (*a*) a general ΔΣ modulator (interpolator structure); (*b*) linear model of the modulator showing injected quantization noise.

Key Point: *To shape the spectrum of a quantization noise source, it is placed within a feedback loop whose response is designed to attenuate the quantization noise in the band of interest.*

To noise-shape the quantization noise in a useful manner, we choose $H(z)$ such that its magnitude is large from 0 to f_0 (i.e., over the frequency band of interest). With such a choice, the signal transfer function, $S_{TF}(z)$, will approximate unity over the frequency band of interest very similarly to an opamp in a unity-gain feedback configuration. Furthermore, the noise transfer function, $N_{TF}(z)$, will approximate zero over the same band. Thus, the quantization noise is reduced over the frequency band of interest while the signal itself is largely unaffected. The high-frequency noise is not reduced by the feedback as there is little loop gain at high frequencies. However, additional post filtering can remove the out-of-band quantization noise with little effect on the desired signal.

Before choosing specific functions for $H(z)$, note that the maximum level of the in-band input signal, $u(n)$, must remain within the maximum levels of the feedback signal, $y(n)$; otherwise the large gain in $H(z)$ will cause the signal $x(n)$ to saturate. For example, if a 1-bit quantizer having output levels of ± 1 is used, the input signal must also remain within ± 1 for frequencies where the gain of $H(z)$ is large. In fact, for many modulators the input signal needs to be significantly smaller than the bounds of the quantizer output levels to keep the modulator *stable*.[3] However, the maximum level of the input signal, $u(n)$, for frequencies where the gain of $H(z)$ is small will not necessarily cause the signal $x(n)$ to saturate. In other words, the maximum level of the out-of-band input signal can be quite a bit larger than the feedback levels (see Problem 18.6).

18.2.2 First-Order Noise Shaping

Key Point: *Using a single integrator inside a $\Delta\Sigma$ modulator feedback loop places a zero at dc in the noise transfer function providing first-order noise shaping for low-pass signals.*

To realize first-order noise shaping, the noise transfer function, $N_{TF}(z)$, should have a zero at dc (i.e., $z = 1$) so that the quantization noise is high-pass filtered. Since the zeros of $N_{TF}(z)$ are equal to the poles of $H(z)$, we can obtain first-order noise shaping by letting $H(z)$ be a discrete-time integrator (i.e., have a pole at $z = 1$).[4] Specifically,

$$H(z) = \frac{1}{z - 1} \tag{18.18}$$

A block diagram for such a choice is shown in Fig. 18.7.

Time Domain View From a time domain point of view, if the feedback is operating correctly and the system is stable, then the signal $x(n)$ is bounded (i.e., $\neq \infty$). Since the integrator has infinite dc gain, the *average value of the discrete-time integrator's input must exactly equal zero* (i.e., average value of $u(n) - y(n)$ equals zero). This result implies that the average value (i.e., dc value) of $u(n)$ must equal the average value (i.e., dc value) of $y(n)$.

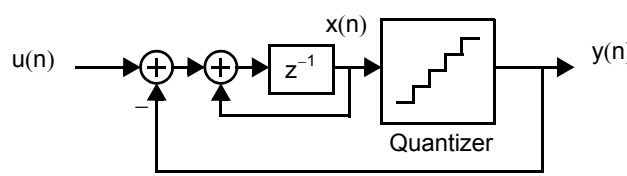


Fig. 18.7 A first-order noise-shaped interpolative modulator.

3. A modulator is defined to be stable if the input to the quantizer does not become so large as to cause the quantizer error to become greater than $\pm\Delta/2$ (which is referred to as overloading the quantizer). See Section 18.7.
4. A continuous-time integrator can also be used, but discrete-time integrators are more popular in integrated realizations as they are less sensitive to sampling jitter and have better distortion characteristics.

Again, the similarity of this configuration and an opamp having unity-gain feedback is emphasized. The open-loop transfer function of an opamp is closely approximated by a first-order integrator having very large gain at low frequencies.

EXAMPLE 18.4

Find the output sequence and state values for a dc input, $u(n)$, of $1/3$ when a two-level quantizer of ± 1.0 is used (threshold at zero) and the initial state for $x(n)$ is 0.1.

Solution

The output sequence and state values are given in Table 18.2.

Note that the average of $y(n)$ exactly equals $1/3$ as expected. However, also note that the output pattern is periodic, which implies that the quantization noise is not random in this example. (However, the result is much more satisfying than applying $1/3$ directly into a 1-bit quantizer using straight oversampling, which would give all 1s as its output.)

Frequency Domain View From a frequency domain view, the signal transfer function, $S_{TF}(z)$, is given by

$$S_{TF}(z) = \frac{Y(z)}{U(z)} = \frac{1/(z-1)}{1+1/(z-1)} = z^{-1} \tag{18.19}$$

and the noise transfer function, $N_{TF}(z)$, is given by

$$N_{TF}(z) = \frac{Y(z)}{E(z)} = \frac{1}{1+1/(z-1)} = (1-z^{-1}) \tag{18.20}$$

We see here that the signal transfer function is simply a delay, while the noise transfer function is a discrete-time differentiator (i.e., a high-pass filter).

To find the magnitude of the noise transfer function, $|N_{TF}(f)|$, we let $z = e^{j\omega T} = e^{j2\pi f/f_s}$ and write the following:

$$\begin{aligned} N_{TF}(f) &= 1 - e^{-j2\pi f/f_s} = \frac{e^{j\pi f/f_s} - e^{-j\pi f/f_s}}{2j} \times 2j \times e^{-j\pi f/f_s} \\ &= \sin\left(\frac{\pi f}{f_s}\right) \times 2j \times e^{-j\pi f/f_s} \end{aligned} \tag{18.21}$$

Table 18.2 First-order modulator example.

n	x(n)	x(n + 1)	y(n)	e(n)
0	0.1	–0.5667	1.0	0.9
1	–0.5667	0.7667	–1.0	–0.4333
2	0.7667	0.1	1.0	0.2333
3	0.1	–0.5667	1.0	0.9
4	–0.5667	0.7667	–1.0	–0.4333
5	. . .	. . .	. . .	. . .

Taking the magnitude of both sides, we have the high-pass function

$$|N_{TF}(f)| = 2\sin\left(\frac{\pi f}{f_s}\right) \tag{18.22}$$

Now the quantization noise power over the frequency band from 0 to f_0 is given by

$$P_e = \int_{-f_0}^{f_0} S_e^2(f)|N_{TF}(f)|^2\,df = \int_{-f_0}^{f_0}\left(\frac{\Delta^2}{12}\right)\frac{1}{f_s}\left[2\sin\left(\frac{\pi f}{f_s}\right)\right]^2 df \tag{18.23}$$

and making the approximation that $f_0 << f_s$ (i.e., $OSR >> 1$), so that we can approximate $\sin((\pi f)/f_s)$ to be $(\pi f)/f_s$, we have

$$P_e \cong \left(\frac{\Delta^2}{12}\right)\left(\frac{\pi^2}{3}\right)\left(\frac{2f_0}{f_s}\right)^3 = \frac{\Delta^2\pi^2}{36}\left(\frac{1}{OSR}\right)^3 \tag{18.24}$$

Assuming the maximum signal power is the same as that obtained before in (18.11), the maximum SNR for this case is given by

$$SQNR_{max} = 10\log\left(\frac{P_s}{P_e}\right) = 10\log\left(\frac{3}{2}2^{2N}\right) + 10\log\left[\frac{3}{\pi^2}(OSR)^3\right] \tag{18.25}$$

or, equivalently,

$$SQNR_{max} = 6.02N + 1.76 - 5.17 + 30\log(OSR) \tag{18.26}$$

Key Point: *First-order noise shaping permits SQNR to increase by 1.5 bits/octave with respect to oversampling ratio, or 9 dB for every doubling of OSR.*

We see here that doubling the OSR gives an SQNR improvement for a first-order modulator of 9 dB or, equivalently, a gain of 1.5 bits/octave. This result should be compared to the 0.5 bits/octave when oversampling with no noise shaping.

18.2.3 Switched-Capacitor Realization of a First-Order A/D Converter

It is possible to realize a first-order modulator using switched-capacitor (SC) techniques. An example of a first-order modulator where a 1-bit quantizer is used in the feedback loop is shown in Fig. 18.8. Here, the ΔΣ modulator consists of both analog and digital circuitry. It should be mentioned that the two input capacitances to the discrete-time integrator in Fig. 18.8 can be combined to one capacitance, as shown in Fig. 18.9 [Boser, 1988]. However, such an approach does not easily allow scaling of the feedback signal relative to the input signal.

18.2.4 Second-Order Noise Shaping

The modulator shown in Fig. 18.10 realizes second-order noise shaping (i.e., the noise transfer function, $N_{TF}(z)$, is a second-order high-pass function). For this modulator, the signal transfer function is given by

$$S_{TF}(f) = z^{-1} \tag{18.27}$$

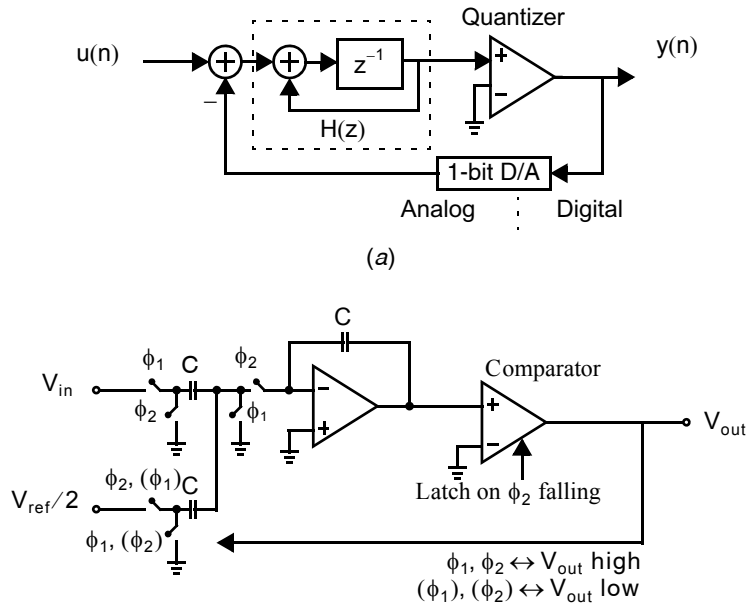


Fig. 18.8 First-order A/D modulator: (*a*) block diagram; (*b*) switched-capacitor implementation.

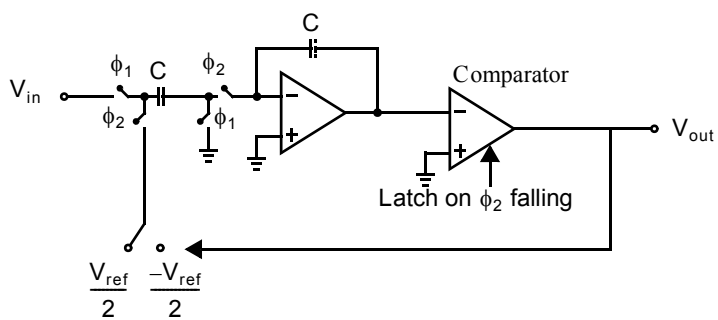


Fig. 18.9 First-order A/D modulator using only one input capacitance to the discrete-time integrator.

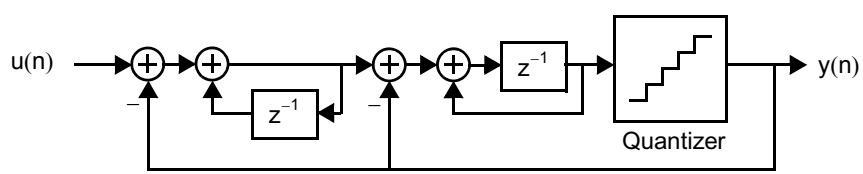


Fig. 18.10 Second-order $\Delta\Sigma$ modulator.

and the noise transfer function is given by

$$N_{TF}(f) = (1 - z^{-1})^2 \tag{18.28}$$

Additionally, the magnitude of the noise transfer function can be shown to be given by

$$|N_{TF}(f)| = \left[2 \sin\left(\frac{\pi f}{f_s}\right)\right]^2 \tag{18.29}$$

resulting in the quantization noise power over the frequency band of interest being given by

$$P_e \cong \frac{\Delta^2 \pi^4}{60}\left(\frac{1}{OSR}\right)^5 \tag{18.30}$$

Again, assuming the maximum signal power is that obtained in (18.11), the maximum SQNR for this case is given by

$$SQNR_{max} = 10 \log\left(\frac{P_s}{P_e}\right) = 10 \log\left(\frac{3}{2} 2^{2N}\right) + 10 \log\left[\frac{5}{\pi^4}(OSR)^5\right] \tag{18.31}$$

or, equivalently,

$$SQNR_{max} = 6.02N + 1.76 - 12.9 + 50 \log(OSR) \tag{18.32}$$

Key Point: *Second-order noise shaping increases SQNR by 15 dB for each doubling of OSR, or equivalently by 1.5 bits/octave.*

We see here that doubling the OSR improves the SQNR for a second-order modulator by 15 dB or, equivalently, a gain of 2.5 bits/octave.

The realization of the second-order modulator using switched-capacitor techniques is left as an exercise for the interested reader.

18.2.5 Noise Transfer-Function Curves

The general shape of zero-, first-, and second-order noise-shaping curves are shown in Fig. 18.11. Note that over the band of interest (i.e., from 0 to f_0), the noise power decreases as the noise-shaping order increases. However, the out-of-band noise increases for the higher-order modulators.

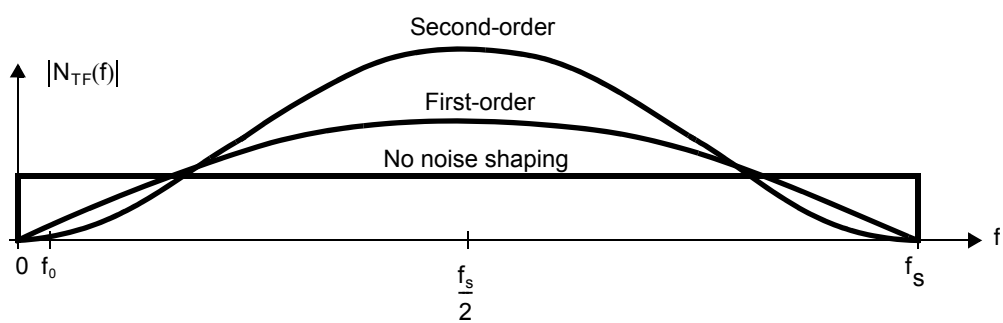


Fig. 18.11 Some different noise-shaping transfer functions.

EXAMPLE 18.5

Given that a 1-bit A/D converter has a 6-dB SQNR, what sample rate is required to obtain a 96-dB SNR (or 16 bits) if $f_0 = 25$ kHz for straight oversampling as well as first- and second-order noise shaping?

Solution

Oversampling with No Noise Shaping From Example 18.3, straight oversampling requires a sampling rate of 54,000 GHz.

First-Order Noise Shaping First-order noise shaping gives 9 dB/octave where 1 octave is doubling the OSR. Since we lose 5 dB, we require 95 dB divided by 9 dB/octave, or 10.56 octaves. Thus, the required sampling rate, f_s, is

$$f_s = 2^{10.56} \times 2f_0 \cong 75\,\text{MHz}$$

This compares very favorably with straight oversampling, though it is still quite high.

Second-Order Noise Shaping Second-order noise shaping gives 15 dB/octave, but loses 13 dB. Thus we required 103 dB divided by 15 dB/octave, resulting in a required sampling rate of only 5.8 MHz. However, this simple calculation does not take into account the reduced input range for a second-order modulator needed for stability.

18.2.6 Quantization Noise Power of 1-Bit Modulators

> ***Key Point:*** *The input signal power is often limited to well below the maximum quantizer swing to ensure stability, resulting in reduced SQNR.*

Assuming the output of a 1-bit modulator is ±1, then one can immediately determine the total power of the output signal, $y(n)$, to be a normalized power of 1 watt. Since $y(n)$ consists of both signal and quantization noise, it is clear that the signal power can never be greater than 1 watt. In fact, as alluded to earlier, the signal level is often limited to well below the ±1 level in higher-order modulators to maintain stability. For example, assuming that the maximum peak signal level is only ±0.25, then the maximum signal power is 62.5 mW (referenced to a 1 Ω load). This would imply a signal power approximately 12 dB less than is predicted by (18.11) and a corresponding reduction in the SQNR estimates that follow from (18.11). Since the signal power plus quantization noise power equals 1 W, in this example the maximum signal power is about 12 dB below the total quantization noise power. Fortunately, as we saw above, the quantization noise power is mostly in a different frequency region than the signal power and can therefore be filtered out. Note, however, that the filter must have a dynamic range capable of accommodating the full power of $y(n)$ at its input. For a ΔΣ A/D converter, the filtering would be done by digital filters following the quantizer.

18.2.7 Error-Feedback Structure

Before leaving this section, it is of interest to look at another structure for realizing a delta-sigma modulator—an error-feedback structure, as shown in Fig. 18.12 [Anastassiou, 1989]. Using a linear model for the quantizer, it is not difficult to show that this error-feedback structure has a signal transfer function, $S_{TF}(z)$, equal to unity while

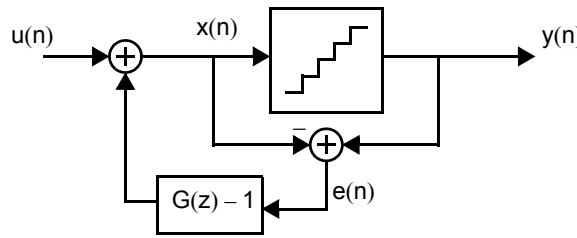


Fig. 18.12 The error-feedback structure of a general $\Delta\Sigma$ modulator.

the noise transfer function, $N_{TF}(z)$, equals $G(z)$. Thus for a first-order modulator, $G(z) = 1 - z^{-1}$, or in other words, the block $(G(z) - 1)$ is simply $-z^{-1}$.

Unfortunately, a slight coefficient error can cause significant noise-shaping degradation with this error-feedback structure. For example, in a first-order case, if the delayed signal becomes $-0.99z^{-1}$ (rather than $-z^{-1}$), then $G(z) = 1 - 0.99z^{-1}$, and the zero is moved off dc. Such a shift of the zero will result in the quantization noise not being fully nulled at dc and therefore would not be suitable for high oversampling ratios. Thus, this structure is not well suited to analog implementations where coefficient mismatch occurs. In contrast, an interpolative structure has the advantage that the zeros of the noise transfer function remain at dc as long as $H(z)$ has infinite gain at dc. This high gain at dc can usually be obtained in analog implementations by using opamps with large open-loop gains and does not rely on component matching. The error feedback structure is discussed here because it is useful for analysis purposes and can work well for fully digital implementations (such as in D/A converters) where no coefficient mismatches occur.

A second-order modulator based on the error-feedback structure is shown in Fig. 18.13. For this case we have

$$G(z) - 1 = z^{-1}(z^{-1} - 2) \tag{18.33}$$

implying that

$$\begin{aligned} G(z) &= 1 - 2z^{-1} + z^{-2} \\ &= (1 - z^{-1})^2 \end{aligned} \tag{18.34}$$

which is identical to (18.28), and we see that second-order noise shaping is obtained. In practical realizations, one might have to take into account additional details such as preventing overflow and reducing complexity. For these more advanced topics, the reader is referred to [Temes, 1996].

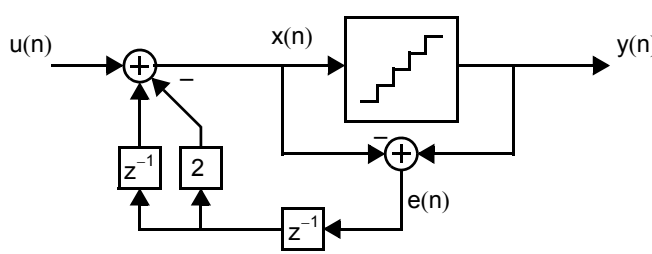


Fig. 18.13 The error-feedback structure of a second-order $\Delta\Sigma$ modulator.

18.3 SYSTEM ARCHITECTURES

In this section, we look at typical system architectures for oversampled A/D and D/A converters.

18.3.1 System Architecture of Delta-Sigma A/D Converters

The system architecture for a typical ΔΣ oversampling A/D converter is shown in Fig. 18.14, and some example signal spectra are shown in Fig. 18.15. In the case of digital audio, various sampling frequencies might be $f_s = 5.6448$ MHz, $f_0 = 44.1$ kHz, which represent an oversampling ratio of 128. Here, the input signal, $x_c(t)$, is sampled and held,[5] resulting in the signal $x_{sh}(t)$. This sampled-and-held signal is then applied to an A/D ΔΣ modulator, which has as its output a 1-bit digital signal $x_{dsm}(n)$. This 1-bit digital signal is assumed to be linearly related to the input signal $x_c(t)$ (accurate to many bits of resolution), although it includes a large amount of out-of-band quantization noise. To remove this out-of-band quantization noise, a digital decimation filter is used as shown. Conceptually, one can think of the decimation process as first reducing the quantization noise through the use of a digital low-pass filter, resulting in the multi-bit signal $x_{lp}(n)$. Note that this low-pass filter will also remove any higher-frequency signal content that was originally on the input signal, $x_c(t)$, and thus also acts as an anti-aliasing filter to limit signals to one-half the final output sampling rate, $2f_0$, as opposed to the anti-aliasing filter at the input, which needed to only limit signals to frequencies less than $f_s/2$. By contrast, in a Nyquist-rate A/D converter, where f_s is only slightly greater than $2f_0$, the analog anti-aliasing filter must have a very sharp cutoff to prevent unwanted signal content (including thermal noise) from aliasing into the band of interest. Hence, oversampling A/D converters require additional digital lowpass filtering, but less analog anti-aliasing filtering—often an desirable trade-off in integrated circuits.

> ***Key Point:*** *Oversampling A/D converters require additional digital low-pass filtering, but less analog anti-aliasing filtering—often an desirable trade-off in integrated circuits.*

Next, $x_{lp}(n)$ is resampled at $2f_0$ to obtain $x_s(n)$ by simply keeping samples at a submultiple of the oversampling rate and throwing away the rest. In Fig. 18.15 an oversampling rate of only 6 is shown for clarity as opposed to the more typical values of 64 or 128 that are used in many commercial applications. This decimation process does not result in any loss of information, since the bandwidth of the original signal was assumed to be f_0. In other words, the signal $x_{lp}(n)$ has redundant information since it is an oversampled signal where all of its spectral information lies well below π, and by throwing away samples, the spectral information is spread over 0 to π. Finally, it should be noted that there is no need to actually create the signal $x_{lp}(n)$, and much digital circuit complexity can be saved by combining the digital low-pass filter with the resampling block to directly

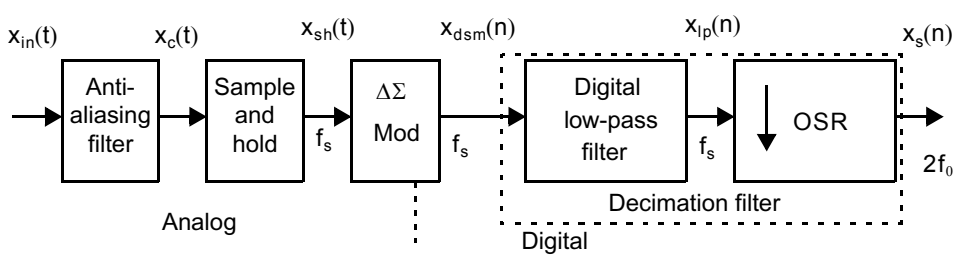


Fig. 18.14 Block diagram of an oversampling A/D converter.

5. This sample-and-hold block is often inherent in the switched-capacitor modulator. Thus, the signal $X_{sh}(t)$ may never physically exist in many realizations.

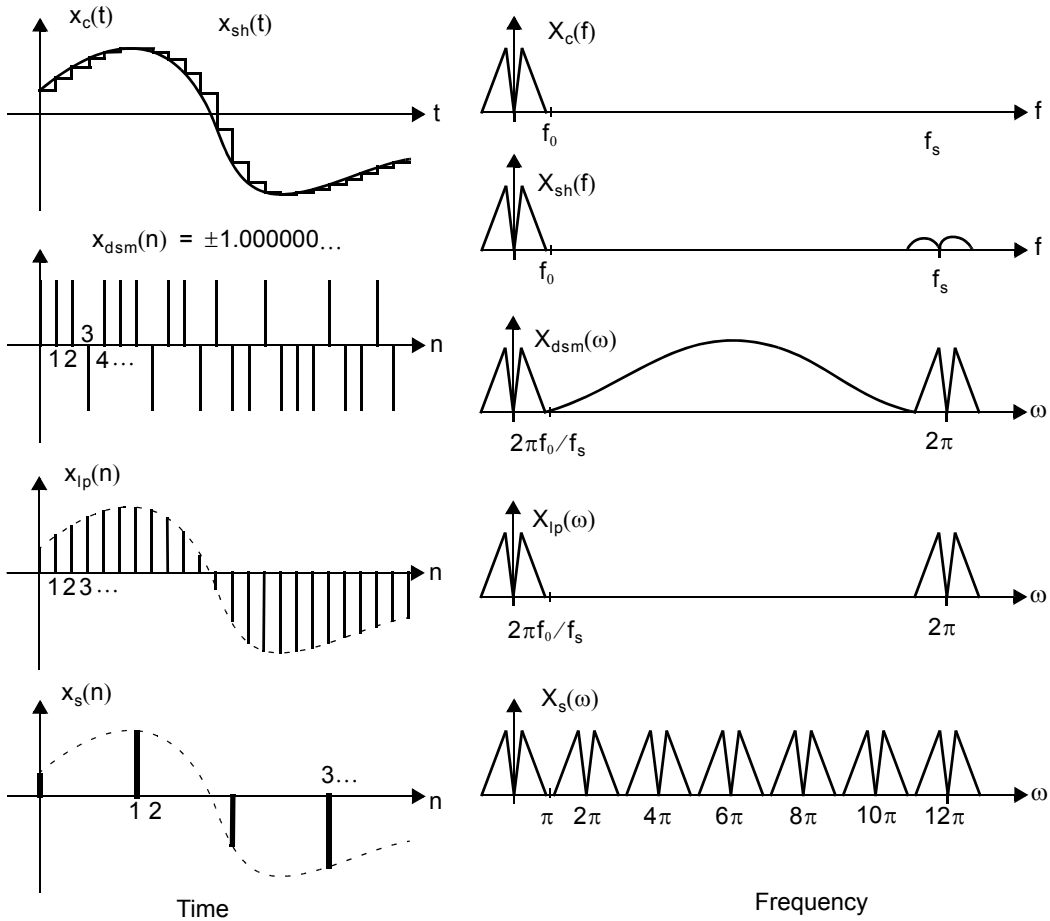


Fig. 18.15 Signals and spectra in an oversampling A/D converter.

produce the downsampled signal $x_s(n)$, as discussed in Section 18.4. The final signal, $x_s(n)$, would typically have 16-bit resolution in digital audio applications.

Key Point: *The linearity of oversampled A/D converters depends more strongly on the linearity of its internal D/A converter than on its internal quantizer.*

It is of interest to look at what element most strongly affects the linearity of this oversampling A/D system. Returning to the ΔΣ modulator, note that an internal 1-bit D/A converter is used whose output signal is combined with the input signal such that, over the frequency band of interest, these two signals are nearly the same. As a result, the overall linearity of this ΔΣ modulator converter depends strongly on the linearity of its internal D/A converter. For example, with a nonlinear internal D/A converter and a slow linear ramp input signal, the low-frequency content of the D/A converter's output would essentially equal that ramp. However, the low-frequency content of the digital input to the D/A converter would be a nonlinear ramp to account for the D/A converter's nonlinearity. Since the remaining digital circuitry is linear, the overall linearity of this oversampling A/D converter is most strongly dependent on realizing a linear D/A converter inside the ΔΣ modulator. In fact, nonlinearities in the internal A/D converter (if it was multi-bit) have only a small effect on the linearity of the overall converter, since the high gain in the feedback loop compensates for that nonlinearity.

18.3.2 System Architecture of Delta-Sigma D/A Converters

Most of the discussion so far has focused on ΔΣ A/D converters, but much of it also applies to ΔΣ D/A converters, with a few qualifications. A high-resolution oversampling D/A converter using a 1-bit converter can be realized as shown in the block diagram in Fig. 18.16. Some illustrative signals that might occur in this system are shown in Fig. 18.17. The digital input signal, $x_s(n)$, is a multi-bit signal and has an equivalent sample rate of $2f_0$, where f_0 is slightly greater than the highest input signal frequency. For example, in a compact disc audio application, a 16-bit input signal is used with a frequency band of interest from 0 to 20 kHz while the sample rate, $2f_0$, is 44.1 kHz. Note, however, that $x_s(n)$ is actually just a series of numbers and thus its shown frequency spectrum has normalized the sample rate to 2π. Next, the input signal is upsampled to an equivalent higher sampling rate, f_s, resulting in the upsampled signal $x_{s2}(n)$. In the example shown, the oversampling rate is only six (i.e., $f_s = 6 \times 2f_0$), while in typical audio applications, f_s is often near a 5-MHz rate (i.e., oversampling rate of 128). However, $x_{s2}(n)$ has large images left in its signal so an interpolation filter is used to create the multi-bit digital signal, $x_{lp}(n)$, by digitally filtering out the images. This interpolation filter is effectively a digital brick-wall-type filter that passes 0 to $(2\pi f_0)/f_s$ and rejects all other signals. The resulting signal, $x_{lp}(n)$, is then applied to a fully digital ΔΣ modulator that produces a 1-bit output signal, $x_{dsm}(n)$, which has a large amount of shaped quantization noise. As discussed earlier, the main reason for going to a 1-bit digital signal is so that a 1-bit D/A converter can now be used to create $x_{da}(t)$, which has excellent linearity properties but still a large amount of out-of-band quantization noise. Finally, the desired output signal, $x_c(t)$, can be obtained by using an analog filter to filter out this out-of-band quantization noise. The analog filter may be a combination of switched-capacitor and continuous-time filtering.

Some points to note here are that while oversampling allows the use of a 1-bit D/A converter, which can have excellent linearity, the use of oversampling also relaxes some of the analog-smoothing filter specifications. For example, if a 16-bit converter was used at the Nyquist rate, $2f_0$, the analog-smoothing filter would have to remove the images in the signal instead of a precise digital interpolation filter removing these images. This specification can be very demanding, as in a digital-audio application, where a near 96-dB attenuation is required at 24 kHz, while up to 20 kHz should pass with unity gain. With the use of an oversampling approach, the digital interpolation filter is faced with this strict specification rather than an analog smoothing filter. In fact, oversampling is often used in audio applications with multi-bit D/A converters just to reduce this analog-smoothing filter's complexity.

Another point to note is that the order of the analog low-pass filter should be at least one order higher than that of the ΔΣ modulator. The reason for this choice is that if the analog filter's order is equal to that of the modulator, the slope of the rising quantization noise will match the filter's falling attenuation, and thus the resulting quantization noise will be approximately a constant spectral density up to one-half the sampling rate (i.e., $f_s/2$). By having a higher-order analog filter, the

Key Point: *Oversampling D/A converters require a very linear but low-resolution internal D/A converter, and an analog smoothing filter whose order should be at least one greater than that of the noise shaping and must have sufficient dynamic range to accommodate both signal and large quantization noise power at its input.*

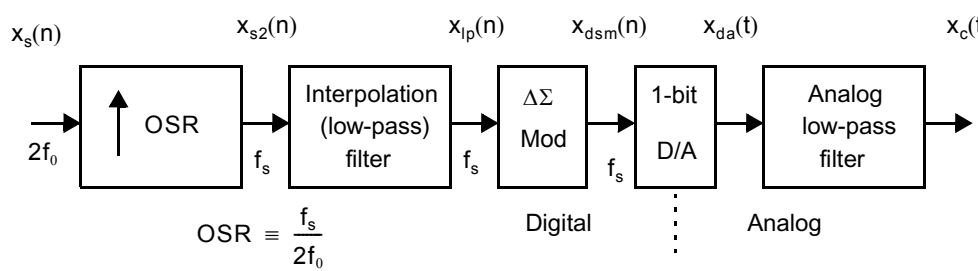


Fig. 18.16 Block diagram of a 1-bit oversampling D/A converter.

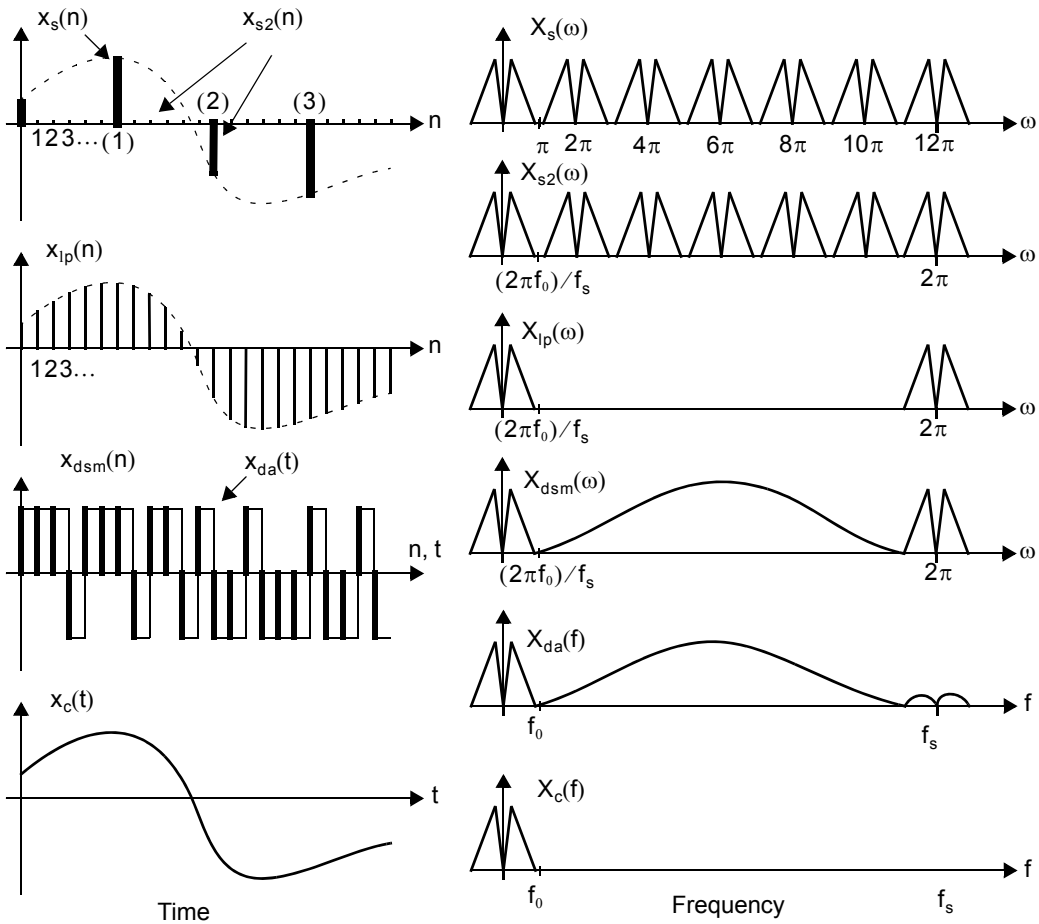


Fig. 18.17 Signals and spectra in an oversampling D/A converter.

spectral density of the output signal's quantization noise will have a bandwidth similar to the analog filter's bandwidth, which is around f_0.

Finally, note that the analog filter must be able to strongly attenuate high-frequency noise, as much of the quantization noise is centered around $f_s/2$, and this analog filter should be linear so it does not modulate the noise back to the frequency band of interest. In many applications, the realization of these filters, especially if they are integrated, is nontrivial.

18.4 DIGITAL DECIMATION FILTERS

There are many techniques for realizing digital decimation filters in oversampling A/D converters. In this section, we discuss two popular approaches—multi-stage and single-stage.

18.4.1 Multi-Stage

One method for realizing decimation filters is to use a multi-stage approach, as shown in Fig. 18.18. Here, the first-stage FIR filter, $T_{sinc}(z)$, removes much of the quantization noise such that its output can be downsampled to four times the Nyquist rate (i.e., $8f_0$). This lower-rate output is applied to the second-stage filter, which may be either an IIR filter, as shown in Fig. 18.18(*a*), or a cascade of FIR filters, as shown in Fig. 18.18(*b*).

Key Point: *In a multi-stage decimation filter, an initial lowpass filter removes much of the quantization noise. Its output is then downsampled so that the remaining filtering can be performed at a lower clock frequency.*

The sinc^{L+1} FIR filter is a cascade of $L + 1$ averaging filters where the transfer function of a single averaging filter, $T_{avg}(z)$, is given by

$$T_{avg}(z) = \frac{Y(z)}{U(z)} = \frac{1}{M}\sum_{i=0}^{M-1} z^{-i} \tag{18.35}$$

and M is the integer ratio of f_s to $8f_0$ (i.e., $8f_0 = f_s/M$). Note that the impulse response of this filter is finite, implying it is an FIR-type filter. In addition, all of its impulse response coefficients are symmetric (in fact, they are all equal), and thus it is also a linear-phase filter.[6] Finally, note that the $1/M$ multiplication term is easily realized by changing the location of the fractional bits when M is chosen to be a power of two.

As an illustration of a series of averaging filters reducing quantization noise, consider an average-of-four filter (i.e., $M = 4$) operating on the output of the 1-bit signal in Example 18.4. Since the output 1-bit sequence is $\{1, 1, -1, 1, 1, -1, 1, 1, -1, \ldots\}$, the first averaged output, $x_{lp1}(n)$, would be given by

$$x_{lp1}(n) = \{0.5, 0.5, 0.0, 0.5, 0.5, 0.0, \ldots\}$$

To obtain more attenuation of the quantization noise, the signal $x_{lp1}(n)$ can also be applied to a running-average-of-four filter, giving $x_{lp2}(n)$ as

$$x_{lp2}(n) = \{0.375, 0.375, 0.25, 0.375, 0.375, 0.25, \ldots\}$$

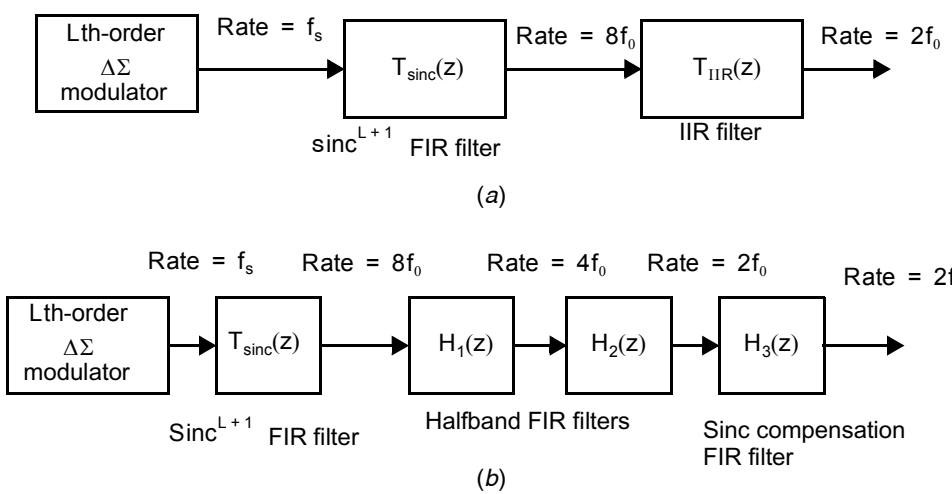


Fig. 18.18 Multi-stage decimation filters: (*a*) sinc followed by an IIR filter; (*b*) sinc followed by halfband filters.

6. A linear-phase filter results in all frequency components being delayed by the same amount and is therefore desirable in applications where phase distortion is unwanted (such as hi-fi audio).

and repeating this process to get $x_{lp3}(n)$ would give

$$x_{lp3}(n) = \{0.344, 0.344, 0.313, 0.344, 0.344, 0.313, \ldots\}$$

Note the convergence of these sequences to a series of all samples equalling $1/3$ as expected.

To show that the frequency response of an averaging filter, $T_{avg}(z)$, has a sinc-type behavior, it is useful to rewrite (18.35) as

$$MY(z) = \left(\sum_{i=0}^{M-1} z^{-i}\right) U(z) = (1 + z^{-1} + z^{-2} + \cdots + z^{-(M-1)})U(z) \tag{18.36}$$

which can also be rewritten as

$$\begin{aligned} MY(z) &= (z^{-1} + z^{-2} + \cdots + z^{-M})U(z) + (1 - z^{-M})U(z) \\ &= Mz^{-1}Y(z) + (1 - z^{-M})U(z) \end{aligned} \tag{18.37}$$

Finally, we group together $Y(z)$ terms and find the transfer function of this averaging filter can also be written in the recursive form as

$$T_{avg}(z) = \frac{Y(z)}{U(z)} = \frac{1}{M}\left(\frac{1 - z^{-M}}{1 - z^{-1}}\right) \tag{18.38}$$

The frequency response for this filter is found by substituting $z = e^{j\omega}$, which results in

$$T_{avg}(e^{j\omega}) = \frac{\operatorname{sinc}\left(\frac{\omega M}{2}\right)}{\operatorname{sinc}\left(\frac{\omega}{2}\right)} \tag{18.39}$$

where $\operatorname{sinc}(x) \equiv \sin(x)/x$.

A cascade of $L + 1$ averaging filters has the response $T_{sinc}(z)$ given by

$$T_{sinc}(z) = \frac{1}{M^{L+1}}\left(\frac{1 - z^{-M}}{1 - z^{-1}}\right)^{L+1} \tag{18.40}$$

The reason for choosing to use $L + 1$ of these averaging filters in cascade is similar to the argument that the order of the analog low-pass filter in an oversampling D/A converter should be higher than the order of the $\Delta\Sigma$ modulator. Specifically, the slope of the attenuation for this low-pass filter should be greater than the rising quantization noise, so that the resulting noise falls off at a relatively low frequency. Otherwise, the noise would be integrated over a very large bandwidth, usually causing excessive total noise.

An efficient way to realize this cascade-of-averaging filters is to write (18.40) as

$$T_{sinc}(z) = \left(\frac{1}{1 - z^{-1}}\right)^{L+1} (1 - z^{M})^{L+1} \frac{1}{M^{L+1}} \tag{18.41}$$

and thus realize it as shown in Fig. 18.19 [Candy, 1992]. A point to note here is that at first glance it appears as though this circuit will not operate properly due to a dc input causing saturation of the discrete-time integrators.

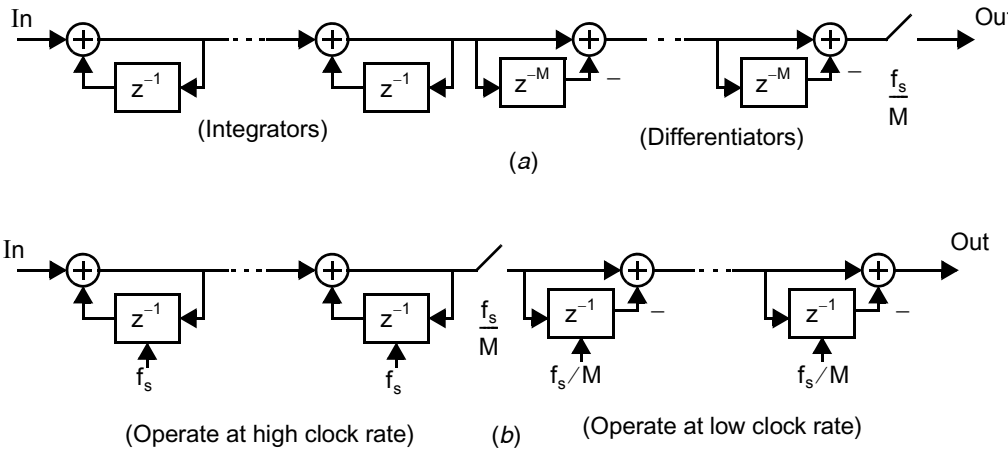


Fig. 18.19 Realizing $T_{sinc}(z)$ as a cascade of integrators and differentiators: (*a*) downsampling performed after all the filtering; (*b*) a more efficient method where downsampling is done before the differentiators.

Fortunately, such a result does not occur when 2's-complement arithmetic is used due to its wrap-around characteristic. Specifically, although the dc levels at the integrator outputs may be incorrect, the differentiators compare the present sample with that of M samples past while rejecting the dc component. Thus, as long as this difference is properly maintained (as it would be using 2's-complement arithmetic) the correct calculation is performed.

Referring once again to Fig. 18.18, the purpose of the filters following the $T_{sinc}(z)$ filter is twofold. One reason for their use is to remove any higher-frequency input signals (in effect, a sharp anti-aliasing filter) before the signal is downsampled to the final Nyquist rate (i.e., $2f_0$). In other words, while the $T_{sinc}(z)$ filter is good at filtering out the quantization noise, it is not sharp enough to act as a reasonable anti-aliasing filter for input signals slightly higher than f_0. A second reason is to compensate for the frequency drop in the passband caused by the $T_{sinc}(z)$ filter. This anti-aliasing and sinc-compensation filter might be realized using a single IIR filter, as shown in Fig. 18.18(*a*). Alternatively, a few halfband FIR filters might be used together with a separate sinc-compensation FIR filter, as shown in Fig. 18.18(*b*). A halfband FIR filter has a passband from 0 to $\pi/2$, while its stopband is from $\pi/2$ to π with every second coefficient being zero [Vaidyanathan, 1993]. Thus, with a high enough filter order, its output can be downsampled by a factor of two. It has also been shown that in some applications, these halfband and sinc-compensation filters can be realized using no general multi-bit multipliers [Saramaki, 1990].

18.4.2 Single Stage

Another approach for realizing decimation filters is to use a relatively high-order FIR filter. For example, in [Dattorro, 1989], a 2048 tap FIR filter was used to decimate 1-bit stereo outputs from two $\Delta\Sigma$ modulators, each having an oversampling ratio of 64. While this FIR order seems high, note that no multi-bit multiplications are needed, since the input signal is simply 1-bit, so all multiplications are trivial. In addition, the output need only be calculated at the Nyquist rate (intermediate samples would be discarded anyway) such that 2048 additions are required during one clock cycle at the Nyquist rate. However, if only one accumulator is used to perform all 2048 additions, the clock rate of that accumulator would

Key Point: *In single-stage decimation, one filter operates on a low-resolution noise-shaped signal providing the decimated output. Because of the high filter order, several time-interleaved filters may be used in parallel.*

be 2048 times the Nyquist rate. For example, if the Nyquist rate is 48 kHz, the single accumulator would have to be clocked at 98.3 MHz. To overcome this high clock rate, 32 separate FIR filters are realized (with shared coefficients) in a time-interleaved fashion, with each FIR having 2048 coefficients and each producing an output at a clock rate of 1.5 kHz. In this way, each of the 32 FIR filters uses a single accumulator operating at 3 MHz (i.e., 2048 times 1.5 kHz). The coefficient ROM values were shared between the FIR filters (as well as in the two stereo channels), and the ROM size can also be halved if coefficients are duplicated, as in the case of linear-phase FIR filtering.

Finally, it is also possible to reduce the number of additions by grouping together input bits. For example, if four input bits are grouped together, then a 16-word ROM lookup table can be used rather than using three additions. With such a grouping of input bits, each of the 32 FIR filters would require only 512 additions.

18.5 HIGHER-ORDER MODULATORS

Key Point: *Noise shaping with order L>2 offers the potential for further improvements in SQNR of 6L+3 dB/octave or L+0.5 bits/octave with respect to oversampling ratio.*

In general, it can be shown that an Lth-order noise-shaping modulator improves the SQNR by $6L + 3$ dB/octave, or equivalently, $L + 0.5$ bits/octave. In this section, we look at two approaches for realizing higher-order modulators—interpolative and MASH. The first approach is typically a single high-order structure with feedback from the quantized signal. The second approach consists of a cascade of lower-order modulators, where the latter modulators are used to cancel the noise errors introduced by the earlier modulators.

18.5.1 Interpolative Architecture

Key Point: *Interpolative higher-order modulators permit the arbitrary placement of noise transfer function zeros, offering improved SQNR performance compared to the placement of all zeros at dc, especially for relatively low OSR. However, it is nontrivial to guarantee stability for such structures, and they are therefore often combined with multi-bit internal quantizers.*

As discussed earlier, when compared to the error-feedback structure, the interpolative structure of Fig. 18.6(*a*) is much better suited to analog implementations of modulators due to its reduced sensitivity. One of the first approaches for realizing higher-order interpolative modulators was presented in [Chao, 1990]. It used a filtering structure very similar to a direct-form filter structure; however, a direct-form-type structure made it sensitive to component variations, which can cause the zeros of the noise transfer function to move off the unit circle. To improve component sensitivity, resonators can be used together with a modified interpolative structure, as shown in Fig. 18.20 [Ferguson, 1991]. Note here that a single 1-bit D/A signal is still used for feedback, and therefore its linearity will

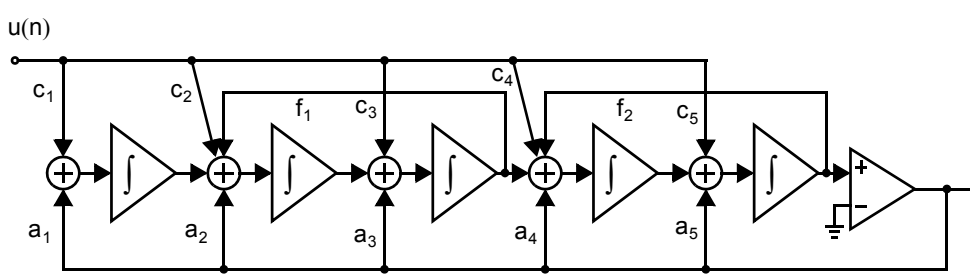


Fig. 18.20 A block diagram of a fifth-order modulator.

be excellent. The resonators in this structure are due to the feedback signals associated with f_1 and f_2, resulting in the placement of zeros in the noise transfer function spread over the frequency-of-interest band. Such an arrangement gives better dynamic range performance than placing all the zeros at dc as we have been assuming thus far.

Unfortunately, it is possible for modulators of order two or more to go unstable, especially when large input signals are present. When they go unstable, they may never return to stability even when the large input signals go away. Guaranteeing stability for an interpolative modulator is nontrivial and is discussed further in Section 18.7. In particular, the use of multi-bit internal quantizers improves stability and is therefore often used in combination with higher-order interpolative architectures.

18.5.2 Multi-Stage Noise Shaping (MASH) Architecture

Another approach for realizing modulators is to use a cascade-type structure where the overall higher-order modulator is constructed using lower-order ones. The advantage of this approach is that since the lower-order modulators are more stable, the overall system should remain stable. Such an arrangement has been called MASH (Multi-stAge noise SHaping) [Matsuya, 1987].

The arrangement for realizing a second-order modulator using two first-order modulators is shown in Fig. 18.21. The basic approach is to pass along the first section's quantization error, $e_1(n)$, to another modulator and combine the outputs of both modulators in such a way that the first section's quantization noise is removed completely. The output is then left with only the second section's quantization noise, which has been filtered twice—once by the second modulator and once by the post digital circuitry. Assuming linear models for the quantizers, straightforward linear analysis for the first stage gives

$$\begin{aligned} X_1(z) &= S_{TF1}(z)U(z) + N_{TF1}(z)E_1(z) \\ &= z^{-1}U(z) + (1 - z^{-1})E_1(z) \end{aligned} \tag{18.42}$$

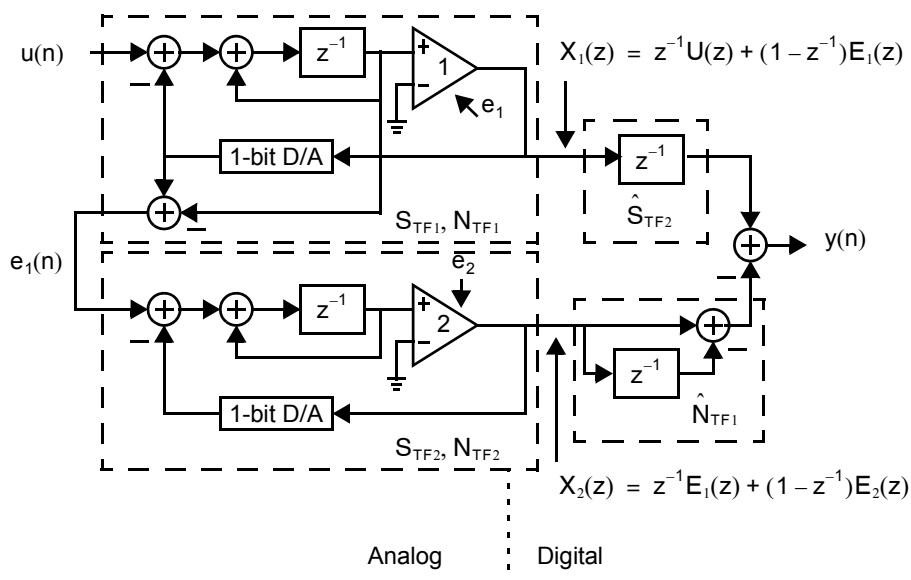


Fig. 18.21 A second-order MASH modulator using two first-order modulators. Note that the output, y(n), is a four-level signal.

and for the second stage

$$\begin{aligned} X_2(z) &= S_{TF2}(z)E_1(z) + N_{TF2}(z)E_2(z) \\ &= z^{-1}E_1(z) + (1 - z^{-1})E_2(z) \end{aligned} \tag{18.43}$$

The MASH output is then formed by passing the two digital outputs x_1 and x_2 through digital filters intended to match S_{TF2} and N_{TF1} respectively.

$$\begin{aligned} Y(z) &= \hat{S}_{TF2}(z)X_1(z) + \hat{N}_{TF1}(z)X_2(z) \\ &= \hat{S}_{TF2}(z)S_{TF1}(z)U(z) + [\hat{S}_{TF2}(z)N_{TF1}(z) + \hat{N}_{TF1}(z)S_{TF2}(z)]E_1(z) \\ &\quad + \hat{N}_{TF1}(z)N_{TF2}(z)E_2(z) \end{aligned} \tag{18.44}$$

If $\hat{N}_{TF1}(z) = N_{TF1}(z)$ and $\hat{S}_{TF2}(z) = S_{TF2}(z)$ then the entire second term disappears resulting in

$$Y(z) = \hat{S}_{TF2}(z)S_{TF1}(z)U(z) + \hat{N}_{TF1}(z)N_{TF2}(z)E_2(z) \tag{18.45}$$

so that the only quantization noise remaining is E_2 shaped by both N_{TF1} and N_{TF2}. In the case of Fig. 18.21, this implies second-order noise shaping using two first-order modulators.

$$Y(z) = z^{-2}U(z) - (1 - z^{-1})^2E_2(z) \tag{18.46}$$

Key Point: *Multi-stage noise shaping (MASH) oversampling converters cascade lower order modulators to realize high-order noise shaping. Since the constituent stages are each first- or second-order, it is easy to ensure stability. However, MASH converters require digital filtering whose coefficients are matched to those of analog circuitry.*

The approach can be generalized and extended to accommodate cascading more than two modulators. Thus, a MASH approach has the advantage that higher-order noise filtering can be achieved using lower-order modulators. The lower-order modulators are much less susceptible to instability as compared to an interpolator structure having a high order with a single feedback. Unfortunately, MASH or cascade approaches are sensitive to finite opamp gain, bandwidth and mismatches causing gain errors. Such errors cause the analog integrators to have finite gain and pole frequencies (see Section 18.7.5), making it difficult to ensure $N_{TF1}(z) = N_{TF1}(z)$ and $\hat{S}_{TF2}(z) = S_{TF2}(z)$ precisely as required for cancellation of the quantization noise $E_1(z)$ in (18.44). In the above example, this would cause first-order noise to leak through from the first modulator and hence reduce dynamic range performance.

In order to mitigate the mismatch problem, the first stage may be chosen to be a higher-order modulator such that any leak-through of its noise has less effect than if this first modulator was first-order. For example, a third-order modulator would be realized using a second-order modulator for the first stage and a first-order modulator for the second stage.

Another approach to combat errors in MASH architectures is to cancel these errors digitally by appropriately modifying the digital filters $\hat{N}_{TF1}$ and $\hat{S}_{TF2}$ in Fig. 18.21. For example, test signals may be injected and later cancelled digitally while using them to precisely identify N_{TF1} and S_{TF2} and match them with $\hat{N}_{TF1}$ and $\hat{S}_{TF2}$ [Kiss, 2000]. Also, it is very important to minimize errors due to input-offset voltages that might occur because of clock feedthrough or opamp input-offset voltages. Typically, in practical realizations additional circuit design techniques will be employed to minimize these effects.

Finally, note that the use of this MASH approach results in the digital output signal, $y(n)$, being a four-level signal (rather than two-level) due to the combination of the original two-level signals. Such a four-level signal would require a linear four-level D/A converter in a D/A application. For A/D applications, it makes the FIR decimation filter slightly more complex.

18.6 BANDPASS OVERSAMPLING CONVERTERS

As we have seen, oversampling converters have the advantage of high linearity through the use of 1-bit conversion and reduced anti-aliasing and smoothing-filter requirements. However, at first glance some signals do not appear to easily lend themselves to oversampling approaches such as modulated radio signals. Such signals have information in only a small amount of bandwidth, say 10 kHz, but have been modulated by some higher-frequency carrier signal, say 1 MHz. For such applications, one can make use of bandpass oversampling converters.

In low-pass oversampling converters, the transfer function $H(z)$ in Fig. 18.6 is chosen such that it has a high gain near dc, and thus quantization noise is small around dc. In a similar way, bandpass oversampling converters are realized by choosing $H(z)$ such that it has a high gain near some frequency value, f_c, [Schreier, 1989]. With such an approach, the quantization noise is small around f_c, and thus most of the quantization noise can be removed through the use of a narrow bandpass filter of total width f_Δ following the modulator. Thus, in the case of a bandpass A/D converter intended for digital radio, post narrowband digital filtering would remove quantization noise as well as adjacent radio channels. In addition, further digital circuitry would also perform the necessary demodulation of the signal.

An important point here is that the oversampling ratio for a bandpass converter is equal to the ratio of the sampling rate, f_s, to two times the width of the narrow-band filter, $2f_\Delta$, and it does not depend on the value of f_c. For example, consider a bandpass oversampling converter with a sampling rate of $f_s = 4$ MHz, which is intended to convert a 10-kHz signal bandwidth centered around 1 MHz (or $f_c = f_s/4$). In this case, $f_\Delta = 10$ kHz, resulting in the oversampling ratio being equal to

$$\text{OSR} = \frac{f_s}{2f_\Delta} = 200 \tag{18.47}$$

To obtain a high gain in $H(z)$ near f_c, a similar approach to the low-pass case is taken. In a first-order low-pass oversampling converter, $H(z)$ is chosen such that it has a pole at dc (i.e., $z = 1$) and such that the noise transfer function, $N_{TF}(z)$, has a zero at dc. In a second-order bandpass oversampling converter with $f_c = f_s/4$, $H(z)$ is chosen such that it has poles at $e^{\pm j\pi/2} = \pm j$. In other words, $H(z)$ is a resonator that has an infinite gain at the frequency $f_s/4$. The zeros of the noise transfer function for this type of second-order oversampling converter are shown in Fig. 18.22, together with a similar low-pass case.

A block diagram for the example bandpass modulator is shown in Fig. 18.23, where one can derive $H(z)$ to be given by

$$H(z) = \frac{z}{z^2 + 1} \tag{18.48}$$

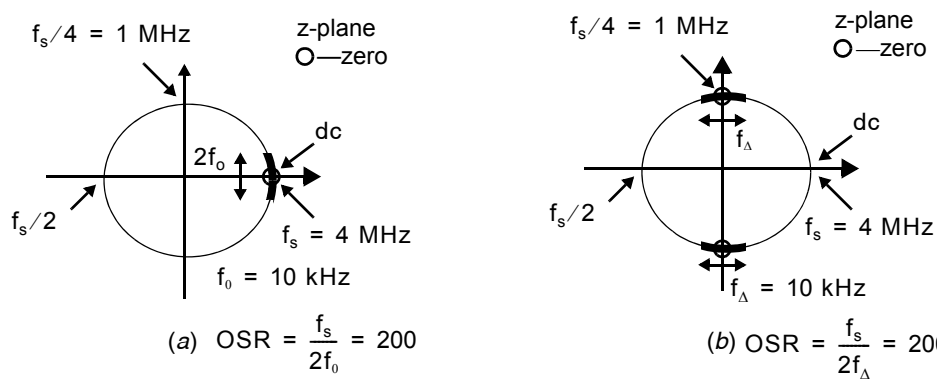


Fig. 18.22 Noise-transfer-function zeros for oversampling converters: (*a*) first-order low-pass; (*b*) second-order bandpass.

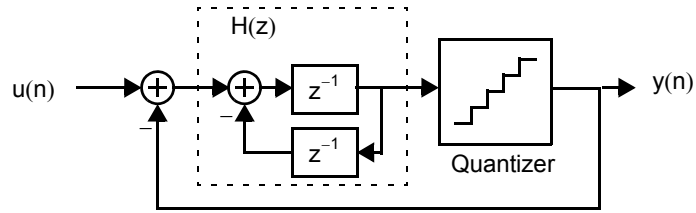


Fig. 18.23 A second-order bandpass oversampling modulator that shapes the quantization noise away from $f_s/4$.

Note that this second-order example has poles at $\pm j$, and therefore the noise transfer function, $N_{TF}(z)$, has only one zero at j and another zero at $-j$. For this reason, the dynamic range increase of a second-order bandpass converter equals that of a first-order low-pass converter that also has only one zero at dc. Specifically, the dynamic range increase of a second-order bandpass converter is 1.5 bits/octave or, equivalently, 9 dB/octave. To obtain the equivalent dynamic range increase of a second-order low-pass converter (i.e., 15 dB/octave), a fourth-order bandpass converter would have to be used. The first design of an integrated bandpass oversampling A/D converter was presented in [Jantzi, 1993].

18.7 PRACTICAL CONSIDERATIONS

18.7.1 Stability

As with any feedback-type system, delta-sigma modulators have the potential to go unstable. A stable modulator is defined here as one in which the input (or inputs) to the quantizer (or quantizers) remains bounded such that the quantization does not become *overloaded.* An overloaded quantizer is one in which the input signal goes beyond the quantizer's normal range, resulting in the quantization error becoming greater than $\pm\Delta/2$.

Unfortunately, the stability of higher-order 1-bit modulators is not well understood as they include a highly nonlinear element (a 1-bit quantizer), resulting in the system being stable for one input but unstable for another. As a general rule of thumb, keeping the peak frequency response gain of the noise-transfer function, $N_{TF}(z)$, less than 1.5 often results in a stable modulator [Chao, 1990].[7] In mathematical terms,

$$\left|N_{TF}(e^{j\omega})\right| \le 1.5 \qquad \text{for } 0 \le \omega \le \pi \tag{18.49}$$

Key Point: *The stability of 1-bit modulators is not well understood, but as a rule of thumb keeping the peak magnitude response of the noise transfer function below 1.5 often results in a stable modulator.*

should be satisfied for a 1-bit quantizer. It should be noted that this stability criterion has little rigorous justification and that there does not presently exist any necessary and sufficient stability criterion for $\Delta\Sigma$ modulators having arbitrary inputs. There is, however, a rigorous 1-norm stability test (which is sufficient but not necessary), but it is often too conservative, as it eliminates many "stable" modulators [Anastassiou, 1989]. Hence, extensive simulations are used to confirm the stability of integrated circuit modulators. For a summary of stability tests for $\Delta\Sigma$ modulators, the reader is referred to [Schreier, 1993].

It should be mentioned here that the stability of a modulator is also related to the maximum input signal level with respect to the 1-bit feedback as well as the poles of the noise transfer function. Specifically, a modulator can be made more stable by placing the poles of the system closer to the noise-transfer-function

7. A figure of 2 was used in the reference but 1.5 appears to be a more practical choice.

zeros. Also, a more stable modulator allows a larger maximum input signal level. However, this stability comes at a dynamic-range penalty since less of the noise power will then occur at out-of-band frequencies, but instead the noise will be greater over the in-band frequencies.

Since the stability issues of higher-order modulators are not well understood (particularly for arbitrary inputs), additional circuitry for detecting instability is often included, such as looking for long strings of 1s or 0s occurring. Alternatively, the signal amplitude at the input of the comparator might be monitored. If predetermined amplitude thresholds are exceeded for a specified number of consecutive clock cycles, then it is assumed that the converter has gone (or is going) unstable. In this case, the circuit is changed to make it more stable. One possibility is to turn on switches across the integrating capacitors of all integrators for a short time. This resets all integrator outputs to zero. A second alternative is to change the modulator into a second- or even first-order modulator temporarily by using only the first one or two integrators and resetting other integrators. Another possibility is to eliminate the comparator temporarily and feed back the input signal to the comparator. This changes the modulator into a stable filter.

Modulators having multi-bit internal quantization exhibit improved stability over their 1-bit counterparts, and are therefore often used in higher order modulators and at low OSR. When used for A/D conversion, they require high-linearity DACs. Fortunately, a number of architectures have been proposed to lessen the linearity requirements on the DACs, as described in Section 18.8.

Key Point: *Multi-bit internal quantization improves stability and therefore is often used in high order modulators with low OSR. In A/D converters, it is combined with linearity enhancement techniques for the feedback DAC.*

18.7.2 Linearity of Two-Level Converters

It was stated earlier that 1-bit D/A converters are inherently linear since they need only produce two output values and two points define a straight line. However, as usual, practical issues limit the linearity of D/A converters—even those having only two output levels. It should be pointed out here that most of the linearity issues discussed below are also applicable to multi-level converters, but we restrict our attention to two-level converters for clarity.

One limitation is when the two output levels somehow become functions of the low-frequency signal they are being used to create. For example, if the two voltage levels are related to power supply voltages, then these power-supply voltages might slightly rise and fall as the circuit draws more or less power. Since typically the amount of power drawn is related to the low-frequency signal being created, distortion will occur. A mechanism whereby this occurs is if a different amount of charge is being taken from the voltage references when a 1 is being output from the 1-bit D/A converter, as opposed to when a 0 is being output from the D/A converter. A somewhat more subtle but similar mechanism can occur due to the clock feedthrough of the input switches. This feedthrough is dependent on the gate voltages driving the input switches and therefore on the power-supply voltages connected to the drivers of the input switches. It is possible that these voltages can change when more 1s than 0s are being output by the D/A converter; having well-regulated power-supply voltages on the drivers for the input switches is very important. A similar argument can be made about any clock jitter in the converter. Thus, for high linearity it is important that the two output levels and associated clock jitter not be a function of the low-frequency input signal.

A more severe linearity limitation occurs if there is memory between output levels. For example, consider the ideal and typical output signals for a two-level D/A converter, shown in Fig. 18.24. The area for each symbol is defined to be the integral of the waveform over that symbol's time period. Notice that for the typical output signal, the area is dependent on the past symbol and that difference is depicted using δ_i. As an illustrative example, consider the average voltage for three periodic patterns corresponding to average voltages of 0, $1/3$, and $-1/3$ when V_1 and V_2 are ± 1 volt.

In the ideal case, both δ_1 and δ_2 are zero as the waveform is memoryless. Thus, the three periodic patterns 0, $1/3$, and $-1/3$ result in the following averages:

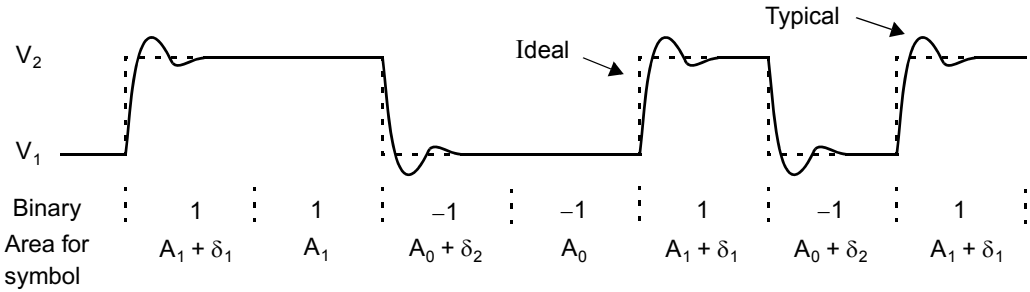


Fig. 18.24 A nonreturn-to-zero (NRZ) 1-bit D/A converter typical output.

$$0: \quad \{1, -1, 1, -1, 1, -1, 1, -1, \ldots\} \rightarrow \overline{v_a(t)} = \frac{A_1 + A_0}{2}$$

$$1/3: \quad \{1, 1, -1, 1, 1, -1, 1, 1, -1, \ldots\} \rightarrow \overline{v_b(t)} = \frac{2A_1 + A_0}{3}$$

$$-1/3: \quad \{-1, -1, 1, -1, -1, 1, -1, -1, 1, \ldots\} \rightarrow \overline{v_c(t)} = \frac{A_1 + 2A_0}{3}$$

By calculating the differences, $\overline{v_b(t)} - \overline{v_a(t)}$ and $\overline{v_a(t)} - \overline{v_c(t)}$, and noting that they are identical, we conclude that $\overline{v_a(t)}$, $\overline{v_b(t)}$, and $\overline{v_c(t)}$ lie along a straight line and therefore this memoryless D/A converter is perfectly linear.

In the typical waveform case, the three periodic patterns result in the following averages:

$$0: \quad \{1, -1, 1, -1, \ldots\} \rightarrow \overline{v_d(t)} = \frac{A_1 + A_0 + \delta_1 + \delta_2}{2} = \overline{v_a(t)} + \frac{\delta_1 + \delta_2}{2}$$

$$1/3: \quad \{1, 1, -1, \ldots\} \rightarrow \overline{v_e(t)} = \frac{2A_1 + A_0 + \delta_1 + \delta_2}{3} = \overline{v_b(t)} + \frac{\delta_1 + \delta_2}{3}$$

$$-1/3: \quad \{-1, -1, 1, \ldots\} \rightarrow \overline{v_f(t)} = \frac{A_1 + 2A_0 + \delta_1 + \delta_2}{3} = \overline{v_c(t)} + \frac{\delta_1 + \delta_2}{3}$$

Now, these three averages do not lie on a straight line except when $\delta_1 = -\delta_2$. Thus, one way to obtain high linearity is to match falling and rising signals—a very difficult task since they are typically realized with different types of devices.

A better way to obtain linearity without matching rising and falling signals of the D/A converter is to use some sort of *memoryless coding sc heme*. For example, consider the return-to-zero coding scheme shown in Fig. 18.25. It is not difficult to see here that since the area of one output bit does not depend on any previous bits, the coding scheme is memoryless and the output signal will remain linear. Naturally, other memoryless schemes can also be used, such as using two levels of opposite signs but ensuring that the signal settles to ground between samples.

It is of interest to see how this memoryless coding is applied in oversampling A/D converters. Presently, most oversampling A/D converters are realized using a switched-capacitor technology for linearity reasons. Switched capacitor circuits naturally realize memoryless levels as long as enough time is left for settling on each clock phase. In other words, switched-capacitor circuits implement memoryless coding since capacitors are charged on one clock phase and discharged on the next. For the same reason, the first stage of filtering in oversampling D/A

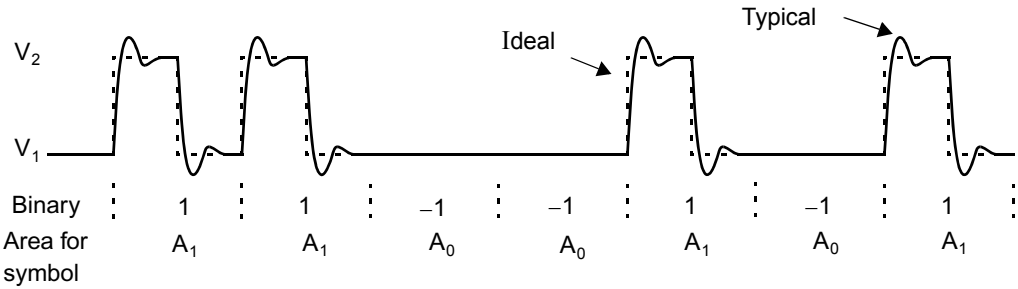


Fig. 18.25 A return-to-zero (RZ) coding scheme, which is one example of a memoryless coding scheme.

converters is often realized using switched-capacitor filters, which is followed by continuous-time filtering. However, some type of memoryless coding should be used when high linearity is desired from a 1-bit D/A output either fed directly to a continuous-time filter in an oversampling D/A or used in the internal operation of a continuous-time oversampling A/D converter.

18.7.3 Idle Tones

Consider the case of applying a dc level of 1/3 to a first-order $\Delta\Sigma$ modulator having a 1-bit quantizer with output levels of ± 1 (as in Example 18.4). With such an input, the output of the modulator will be a periodic sequence of two 1s and one –1. In other words,

$$y(n) = \{1, 1, -1, 1, 1, -1, 1, 1, \ldots\} \tag{18.50}$$

This output periodic pattern is three cycles long and its power is concentrated at dc and $f_s/3$. Although the output pattern is periodic, the post low-pass filter will eliminate the high-frequency content such that only the dc level of 1/3 remains.

Now consider applying a dc level of $(1/3 + 1/24) = 3/8$ to the same modulator. For this case, the output sequence becomes

$$y(n) = \{1, 1, -1, 1, 1, -1, 1, 1, -1, 1, 1, -1, 1, 1, 1, -1, 1, 1, -1, \ldots\} \tag{18.51}$$

The period of this output pattern is now 16 cycles long and has some power at $f_s/16$. With an oversampling ratio of eight (i.e., $f_0 = f_s/16$), the post low-pass filter will not attenuate the signal power at $f_s/16$ since that frequency is just within the frequency band of interest. In other words, a dc level of $3/8$ into this modulator will produce the correct dc output signal but have a superimposed $f_s/16$ signal on it even when using a brick-wall type low-pass filter.

While the preceding example shows a tone occurring at $f_s/16$, it is not difficult to find other cases that would produce frequency tones at much lower frequencies, say $f_s/256$. Such low-frequency tones *will not be filtered out by the next stage low-pass filter* and can lead to annoying tones in the audible range.[8] In addition, although a first-order modulator was used in the above example, it has been shown that annoying audible tones exist even in higher-order modulators [Norsworthy, 1993]. Finally, it should be mentioned that these tones might not lie at a single frequency but instead be short-term periodic patterns. In other words, a tone appearing near 1 kHz might actually be a signal varying between 900 Hz and 1.1 kHz in a random-like fashion.

8. Human hearing can detect tones buried beneath white noise.

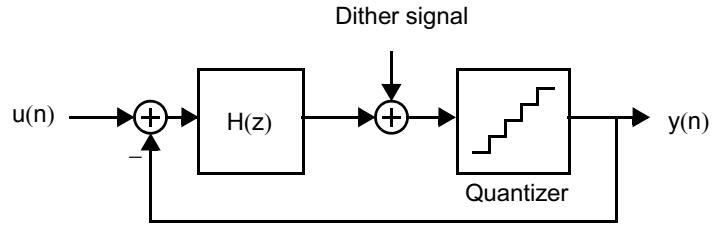


Fig. 18.26 Adding dithering to a delta-sigma modulator. Note that the dithered signal is also noise shaped.

18.7.4 Dithering

One way to reduce the amount of idle tones in modulators is through the use of dithering. The term dithering here refers to the act of introducing some random (or pseudo-random) signal into a modulator.

Key Point: *The nonlinear dynamics of 1-bit oversampling modulators can result in idle tones within the signal band when applying dc or slowly-varying inputs. An additive pseudo-random dither signal can break up the tones but raises the noise floor.*

Assuming the random signal to be introduced has a white-noise type spectrum, then the most suitable place to add the dithering signal is just before the quantizer, as shown in Fig. 18.26. The reason for adding it at this location is so that the dithering signal becomes noise shaped in the same manner as the quantization noise, and thus a large amount of dithering can be added. Typically, the dithering signal is realized using some sort of pseudo-random number generator with only a few bits of resolution, but its total noise power is comparable to the quantization noise power. Thus, the use of dithering in a D/A converter would require a small additional digital adder, while a small D/A converter would be needed in an oversampling A/D converter.

It should be noted that the use of dither to reduce idle tones is not an attempt to add noise to mask out the tones but instead breaks up the tones so that they never occur. However, since the noise power of the dithering signal is similar to the quantization noise power, the use of dithering adds about 3 dB extra in-band noise and often requires rechecking the modulator's stability.

18.7.5 Opamp Gain

A switched-capacitor integrator with a finite opamp gain, A, results in the integrator having a transfer function (see Problem 14.6) of

$$\frac{V_o(z)}{V_i(z)} = -\left(\frac{C_1}{C_2}\right)\left(\frac{1}{z\left(1+\dfrac{C_1}{C_2 A}\right)-1}\right) \tag{18.52}$$

Assuming $C_2 \approx C_1$ as is typical in oversampled converters, the pole has moved to the left of $z = 1$ by an amount $1/A$. With this knowledge, we can determine the new NTF zeros due to finite opamp gain by substituting $z - 1$ with $z - (1 - 1/A)$. In other words, we substitute all ideal integrators with damped integrators and recalculate the transfer functions. Such an approach results in the NTF zeros, which were located at $z = 1$ for a low-pass design, to be moved to $z = (1 - 1/A)$. Thus, the shaped quantization noise does not drop to zero at dc but instead levels off near dc. Such a zero near $z = 1$ results in the 3-dB break frequency being approximately equal to $1/A$ rad/sample. Now, we note that the frequency band of interest, f_0, should be greater than $1/A$ rad/sample since the shaped quantization noise is flat below that level. In other words, if f_0 is below where the quantization noise flattens

out, then we are not obtaining any further *noise-shaping* benefits. Thus, any further doubling of the oversampling ratio will only improve the SNR by 3 dB/octave. Finally, we can write the following requirement:

$$\frac{f_0}{f_s} > \frac{1/A}{2\pi} \tag{18.53}$$

and recalling that $OSR = f_s/(2f_0)$, we can rearrange this to

$$A > \frac{OSR}{\pi} \tag{18.54}$$

Of course, some approximations have been made here, such as having the two capacitors of the integrator equal (that is $C_2 \approx C_1$) and having the 3-dB break point being sharp, as well as allowing noise to be flat from dc to f_0 (rather than shaped). In summary, designers will typically ensure that the opamp gain is at least twice the oversampling ratio, which is not usually a difficult requirement.

In addition, the above analysis only applies to modulators having a single D/A feedback and does not apply to MASH or cascade modulators, where larger opamp gains are often required to aid matching between digital and analog signal paths.

Key Point: *Opamps in analog oversampling modulators should have a dc gain at least twice the oversampling ratio, which is not usually a difficult requirement. An exception is MASH modulators where higher gains may be required to ensure digital and analog transfer functions are matched.*

18.8 MULTI-BIT OVERSAMPLING CONVERTERS

Although 1-bit oversampling converters have the advantage that they can realize highly linear data conversion, they also have some disadvantages. For example, 1-bit oversampling modulators are prone to instability due to the high degree of nonlinearity in the feedback. Another disadvantage is the existence of idle tones, as previously discussed. Additionally, in oversampling D/A converters, the use of a 1-bit D/A converter results in a large amount of out-of-band quantization noise, which must be significantly attenuated using analog circuitry. Such a task often requires relatively high-order analog filtering (recall that the low-pass filter should be at least one order higher than the modulator order). Also, the dynamic range of the analog filter may have to be significantly higher than the converter's dynamic range since the filter's input must cope with power of both the quantization noise and signal itself. Since the quantization noise may be up to 15 dB larger in level than the signal level, such an extension in the dynamic range makes the realization of analog filters for oversampling D/A converters nontrivial. The use of a multi-bit D/A converter can significantly reduce this large amount of quantization noise, but care must be taken to ensure that the multi-bit converter remains linear. Typically, a multi-bit oversampling system would be similar to that shown in Fig. 18.16, except that an M-bit quantizer would be used in the digital modulator, and its output would drive an M-bit D/A converter. This section will briefly discuss some multi-bit oversampling converter architectures that employ methods so that the overall system remains highly linear.

18.8.1 Dynamic Element Matching

One approach for realizing a multi-bit D/A converter intended for oversampling systems is that of dynamic element matching [Carley, 1989]. With this approach, a thermometer-code D/A converter is realized, and randomization is used to effectively spread the nonlinearity over the whole frequency range. For example, a 3-bit D/A converter using dynamic element matching is shown in Fig. 18.27. Here, the 8-line randomizer makes use of

Key Point: *Dynamic element matching randomizes the use of analog elements in a D/A to eliminate the tones that would otherwise result from mismatches, making instead spectrally-shaped noise. It enables the use of multi-bit D/As in highly linear oversampling converters.*

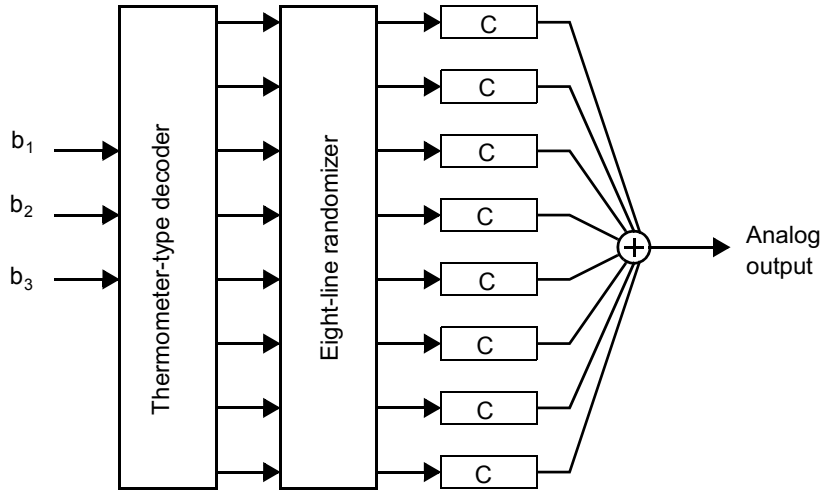


Fig. 18.27 Dynamic element matching 3-bit D/A converter.

a pseudo-random code to distribute the thermometer lines equally to each of the unit capacitances, **C**. Using the simplest randomization approaches, the mismatch "noise" from this multi-bit converter is not shaped by any feedback loop and therefore appears more white than high-pass filtered [Carley, 1989].

Dynamic element matching has been extended to include noise shaping the nonlinearity introduced by the D/A [Leung, 1992; Galton, 1997]. These approaches enable much lower oversampling ratios and can be used while maintaining high linearity with a reasonably coarse DAC mismatch. They are currently the most popular approach to mitigating D/A nonlinearity errors in oversampling converters. Some examples are [Chen, 1995; Adams, 1995; Baird, 1995].

18.8.2 Dynamically Matched Current Source D/A Converters

Another method of maintaining high linearity in an oversampling D/A system is through the use of a highly linear D/A converter operating at the fast oversampling rate. Such high linearity has been achieved using dynamically matched current sources similar to those described in Section 16.3.3 [Schouwenaars, 1991].

18.8.3 Digital Calibration A/D Converter

One way to make use of a multi-bit D/A converter in an oversampling A/D modulator is to have a calibration cycle where digital circuitry is used to model the static nonlinearity of the D/A converter, as shown in Fig. 18.28 [Larson, 1988]. With this approach, the multi-bit digital signal $x_2(n)$ creates 2^N output levels at $x_1(n)$, which are consistent in time but not necessarily linearly related to $x_2(n)$. However, the feedback loop of the modulator forces the in-band frequency of $x_1(n)$ to very closely equal $u(n)$, and thus $x_1(n)$ is still linearly related to $u(n)$. Through the use of the digital correction circuitry, the signal $y(n)$ also equals $x_1(n)$ if its nonlinearity equals that of the D/A converter. To realize this digital correction circuitry, a 2^N word look-up RAM is used where the size of each RAM word is at least equal to the desired linearity. For example, if a 4-bit D/A converter is used and 18 bits of linearity is desired, then the digital correction circuitry consists of a 16-word RAM, each of length 18 bits or greater (or equivalently, a 288-bit RAM). Furthermore, this RAM size can be reduced by digitizing only the difference between the desired linearity and the expected linearity of the D/A converter [Sarhang-Nejad, 1993]. For example, if the D/A converter is

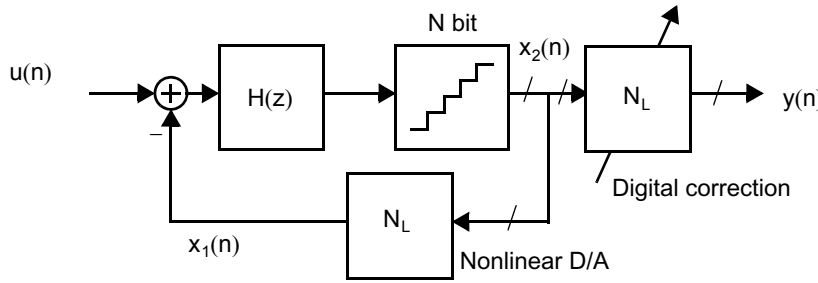


Fig. 18.28 Digitally corrected multi-bit A/D modulator.

expected to have a linearity of at least 9 bits, then the RAM need only have a word length of 10 bits for 18-bit overall linearity.

The calibration of the digital correction circuitry can be accomplished by reconfiguring the $\Delta\Sigma$ A/D modulator to a single-bit system. Next, the input to this 1-bit converter is one of the 4-bit D/A converter levels to be calibrated. In other words, the multi-bit converter's input is held constant while its output is applied to the 1-bit A/D modulator for many clock cycles. The resulting 1-bit signal is digitally low-pass filtered and gives the desired digital equivalent value of that particular dc level for the D/A converter. This procedure is repeated 16 times to find all the 16 words of RAM.

Finally, it should be mentioned that a similar approach can be used to improve the linearity of a multi-bit oversampling D/A converter, except in this case the digital correction circuit would be in the feedback portion of the digital modulator.

18.8.4 A/D with Both Multi-Bit and Single-Bit Feedback

Another very interesting alternative architecture uses both multi-bit feedback and single-bit feedback [Hairapetian, 1994]. A third-order example of this A/D is shown in Fig. 18.29. If one assumes that the errors due to the quantization of the 1-bit A/D and 5-bit A/D are Q_1, and Q_5, respectively, and also that the 5-bit D/A injects errors due to its nonlinearity given by Q_d, then by applying a linear analysis to the system of Fig. 18.29 it can be shown that

$$\begin{aligned} U_s(z) &= U(z)z^{-1} + Q_1(z)\frac{(1-z^{-1})^2}{1-0.5z^{-1}} - Q_5(z)\frac{z^{-1}(1-z^{-1})^{-2}}{1-0.5z^{-1}} \\ &\quad - Q_d(z)\frac{z^{-1}(1-z^{-1})^{-2}}{1-0.5z^{-1}} \end{aligned} \tag{18.55}$$

After digital signal processing, the digital output signal, $Y(z)$, is given by

$$Y(z) = U(z)z^{-1} + 2Q_5(z)(1-z^{-1})^3 - 2Q_d(z)z^{-1}(1-z^{-1})^2 \tag{18.56}$$

It is seen that, based on the assumption of perfect integrators, the quantization noise due to the 1-bit A/D is cancelled exactly, the quantization noise due to the 5-bit A/D undergoes third-order noise shaping, and the noise due to the nonlinearity of the 5-bit D/A converter undergoes second-order shaping. Of course, if the integrators are not perfect, then the cancellation of the quantization noise of the 1-bit A/D will not be perfect, but any errors in the cancellation still undergo second-order noise shaping. It can be shown that the stability characteristics of this system are very similar to those of a second-order $\Delta\Sigma$ A/D converter. It should also be noted that the digital signal processing required does not require any multiplications, but that multi-bit signals must be processed by the following decimator, which complicates its realization.

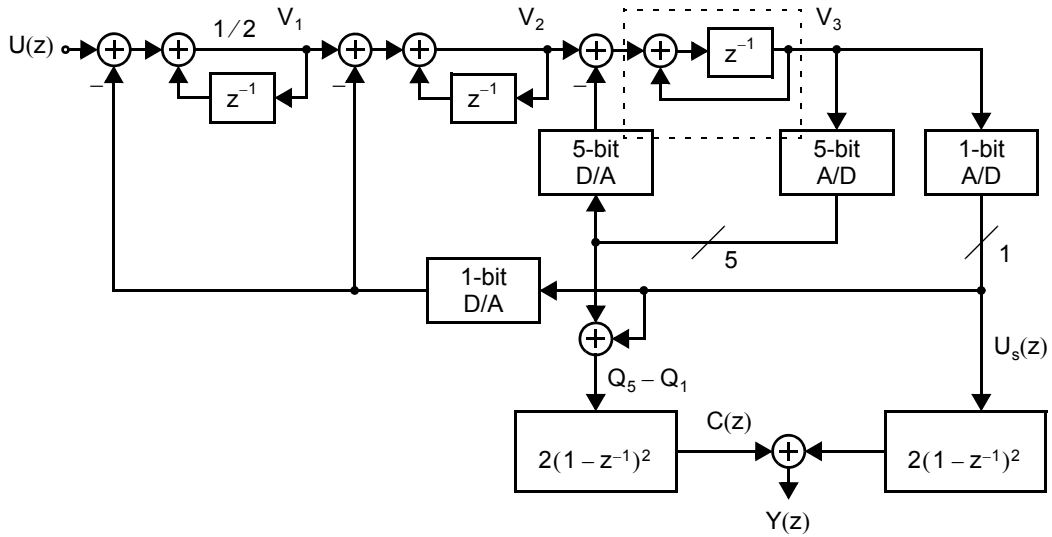


Fig. 18.29 An A/D architecture that uses both multi-bit and single-bit feedback.

18.9 THIRD-ORDER A/D DESIGN EXAMPLE

In this section, a design example is presented for a third-order oversampled switched-capacitor A/D converter. In this example, all the zeros of the noise transfer function (NTF) are placed at $z = 1$ (i.e., dc) so that the converter could be used for various oversampling ratios. In other words, the zeros are not spread over the frequency band of interest, as that would restrict the converter's usefulness to a particular oversampling ratio.

Since all the zeros are assumed to be at $z = 1$, the NTF has the following form:

$$\mathrm{NTF}(z) = \frac{(z-1)^3}{D(z)} \tag{18.57}$$

To find the poles of the system, given by the roots of $D(z)$, we need to ensure modulator stability while shaping as much quantization noise away from dc as possible. To ensure modulator stability, recall that a heuristic approach is to restrict the peak NTF magnitude response to less than 1.5. However, to shape the quantization noise away from dc, we wish to make the NTF as large as possible outside the frequency band of interest. Thus, the best NTF choice would be one which has a flat gain of 1.4 at high frequencies. Such a high-pass filter is obtained (i.e., no peaking in the NTF) when the poles are placed in a Butterworth configuration, as shown in Fig. 18.30. Thus, the

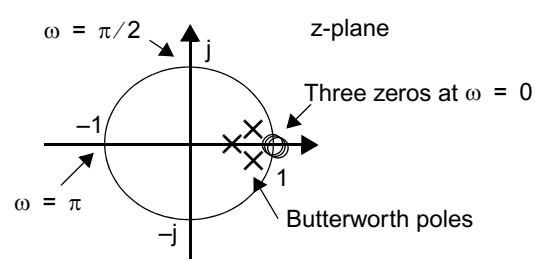


Fig. 18.30 Pole and zero locations for the third-order NTF.

denominator is found using digital filter software to design a third-order high-pass digital filter where the passband is adjusted until the peak NTF is less than 1.4 (i.e., $|\mathrm{NTF}(e^{j\omega})| < 1.4$). Specifically, with a passband edge at $f_s/20$, a third-order Butterworth high-pass filter has a peak gain equal to 1.37, and thus the $\mathrm{NTF}(z)$ is given by

$$\mathrm{NTF}(z) = \frac{(z-1)^3}{z^3 - 2.3741z^2 + 1.9294z - 0.5321} \tag{18.58}$$

By rearranging (18.16), we can find $\mathrm{H}(z)$ in term of $\mathrm{NTF}(z)$, resulting in

$$\mathrm{H}(z) = \frac{1 - \mathrm{NTF}(z)}{\mathrm{NTF}(z)} \tag{18.59}$$

which, for the function given in (18.58), results in

$$\mathrm{H}(z) = \frac{0.6259z^2 - 1.0706z + 0.4679}{(z-1)^3} \tag{18.60}$$

Next, a suitable implementation structure is chosen. In this case, a cascade-of-integrators structure was used, as shown in Fig. 18.31. The α_i coefficients are included for dynamic-range scaling and are initially set to $\alpha_2 = \alpha_3 = 1$, while the last term, α_1, is initially set equal to β_1. By initially setting $\alpha_1 = \beta_1$, we are allowing the input signal to have a power level similar to that of the feedback signal, $\mathrm{y}(n)$. In other words, if α_1 were initially set equal to one and β_1 were quite small, then the circuit would initially be stable for only small input-signal levels.

Coefficient values for β_i are then found by deriving the transfer function from the 1-bit D/A output to V_3 and equating that function to $-\mathrm{H}(z)$ in (18.60). Specifically, assuming $\alpha_2 = \alpha_3 = 1$, we find the following $\mathrm{H}(z)$:

$$\mathrm{H}(z) = \frac{z^2(\beta_1 + \beta_2 + \beta_3) - z(\beta_2 + 2\beta_3) + \beta_3}{(z-1)^3} \tag{18.61}$$

Equating (18.61) with (18.60), we find an initial set of coefficients to be

$$\begin{aligned} \alpha_1 = 0.0232, \quad \alpha_2 = 1.0, \quad \alpha_3 = 1.0 \\ \beta_1 = 0.0232, \quad \beta_2 = 0.1348, \quad \beta_3 = 0.4679 \end{aligned} \tag{18.62}$$

It is then necessary to perform dynamic-range scaling. Dynamic scaling is necessary to ensure that all nodes have approximately the same power level, so that all nodes will clip near the same level, and there will be no unnecessarily large noise gains from nodes with small signal levels. Here, dynamic-range scaling was accomplished by simulating the system in Fig. 18.31 with the coefficients just given and an input sinusoidal wave with a peak value of 0.7 at a frequency of $\pi/256$ rad/sample. Following the simulation, the maximum values at nodes $\mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_3$ were found to be 0.1256, 0.5108, and 1.004, respectively. To increase the output level of a node by a

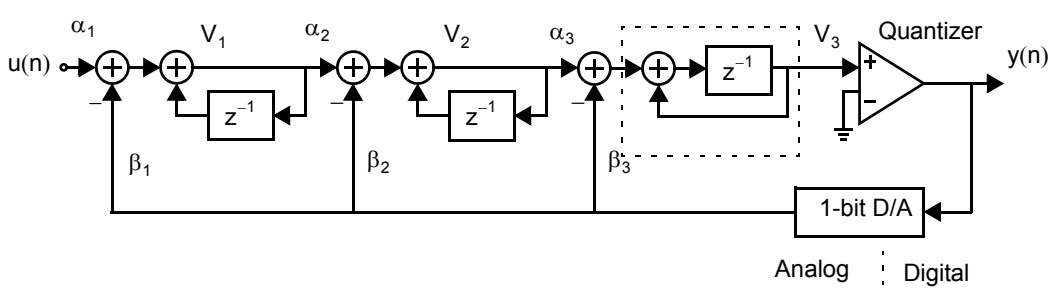


Fig. 18.31 The cascade-of-integrators structure used to realize the third-order modulator. Note that α_i coefficients are used for dynamic-range scaling and are initially set to $\alpha_1 = \beta_1$ and $\alpha_2 = \alpha_3 = 1$.

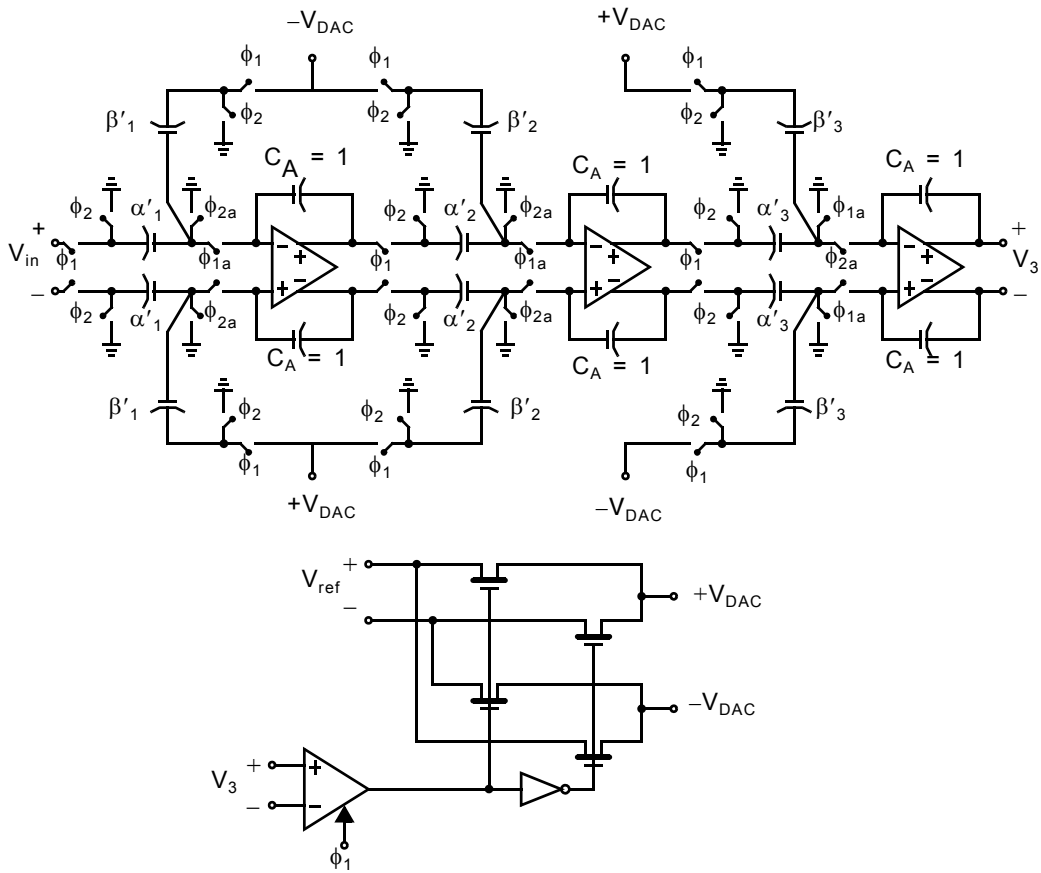


Fig. 18.32 A switched-capacitor realization of the third-order modulator example.

factor k, the input branches should be multiplied by k, while the output branches should be divided by k. For example, the maximum value of node V_1 is increased to unity by multiplying both α_1 and β_1 by $1/0.1256$ and dividing α_2 by $1/0.1256$. Similarly, node V_2 is scaled by dividing $0.1256\alpha_2$ and β_2 by 0.5108 and multiplying α_3 by 0.5108. Node V_3 was left unchanged, as its maximum value was already near unity. Such a procedure results in the following final values for the coefficients.

$$\begin{aligned} \alpha'_1 &= 0.1847, \quad \alpha'_2 = 0.2459, \quad \alpha'_3 = 0.5108 \\ \beta'_1 &= 0.1847, \quad \beta'_2 = 0.2639, \quad \beta'_3 = 0.4679 \end{aligned} \tag{18.63}$$

Finally, the block diagram of Fig. 18.31 can be realized by the switched-capacitor circuit shown in Fig. 18.32 to obtain the final oversampled A/D converter.

18.10 KEY POINTS

- Oversampling converters relax the requirements placed on the analog circuitry at the expense of more complicated digital circuitry. [p. 696]
- Quantization with a step size "LSB" can be modeled as an additive quantization noise uniformly distributed between –LSB/2 and +LSB/2 with a white power spectrum and total power of LSB2/12. [p. 697]

- Oversampling a signal at OSR times the Nyquist rate, without spectrally shaping the quantization noise, results in a SNR enhancement of 10 log(OSR) dB. [p. 700]
- One-bit D/A converters are inherently linear. Hence, when combined with a high oversampling ratio, they can provide a good signal to quantization noise and distortion ratio. [p. 702]
- To shape the spectrum of a quantization noise source, it is placed within a feedback loop whose response is designed to attenuate the quantization noise in the band of interest. [p. 704]
- Using a single integrator inside a DS modulator feedback loop places a zero at dc in the noise transfer function providing first-order noise shaping for low-pass signals. [p. 704]
- First-order noise shaping permits SQNR to increase by 1.5 bits/octave with respect to oversampling ratio, or 9 dB for every doubling of OSR. [p. 706]
- Second-order noise shaping increases SQNR by 15 dB for each doubling of OSR, or equivalently by 1.5 bits/octave. [p. 708]
- The input signal power is often limited to well below the maximum quantizer swing to ensure stability, resulting in reduced SQNR. [p. 709]
- Oversampling A/D converters require additional digital low-pass filtering, but less analog anti-aliasing filtering—often an desirable trade-off in integrated circuits. [p. 711]
- The linearity of oversampled A/D converters depends more strongly on the linearity of its internal D/A converter than on its internal quantizer. [p. 712]
- Oversampling D/A converters require a very linear but low-resolution internal D/A converter, and an analog smoothing filter whose order should be at least one greater than that of the noise shaping and must have sufficient dynamic range to accommodate both signal and large quantization noise power at its input. [p. 713]
- In a multi-stage decimation filter, an initial lowpass filter removes much of the quantization noise. Its output is then downsampled so that the remaining filtering can be performed at a lower clock frequency. [p. 715]
- In single-stage decimation, one filter operates on a low-resolution noise-shaped signal providing the decimated output. Because of the high filter order, several time-interleaved filters may be used in parallel. [p. 717]
- Noise shaping with order $L>2$ offers the potential for further improvements in SQNR of $6L+3$ dB/octave or $L+0.5$ bits/octave with respect to oversampling ratio. [p. 718]
- Interpolative higher-order modulators permit the arbitrary placement of noise transfer function zeros, offering improved SQNR performance compared to the placement of all zeros at dc, especially for relatively low OSR. However, it is nontrivial to guarantee stability for such structures, and they are therefore often combined with multi-bit internal quantizers. [p. 718]
- Multi-stage noise shaping (MASH) oversampling converters cascade lower order modulators to realize high-order noise shaping. Since the constituent stages are each first- or second-order, it is easy to ensure stability. However, MASH converters require digital filtering whose coefficients are matched to those of analog circuitry. [p. 720]
- The stability of 1-bit modulators is not well understood, but as a rule of thumb keeping the peak magnitude response of the noise transfer function below 1.5 often results in a stable modulator. [p. 722]
- Multi-bit internal quantization improves stability and therefore is often used in high order modulators with low OSR. In A/D converters, it is combined with linearity enhancement techniques for the feedback DAC. [p. 723]
- The nonlinear dynamics of 1-bit oversampling modulators can result in idle tones within the signal band when applying dc or slowly-varying inputs. An additive pseudo-random dither signal can break up the tones but raises the noise floor. [p. 726]
- Opamps in analog oversampling modulators should have a dc gain at least twice the oversampling ratio, which is not usually a difficult requirement. An exception is MASH modulators where higher gains may be required to ensure digital and analog transfer functions are matched. [p. 727]

- Dynamic element matching randomizes the use of analog elements in a D/A to eliminate the tones that would otherwise result from mismatches, making instead spectrally-shaped noise. It enables the use of multi-bit D/As in highly linear oversampling converters. [p. 727]

18.11 REFERENCES

R. W. Adams and T. W. Kwan. "Data-Directed Scrambler for Multi-Bit Noise Shaping D/A Converters," U.S. patent no. 5,404,142, April 1995.

D. Anastassiou. "Error Diffusion Coding for A/D Conversion," *IEEE Trans. on Circuits and Systems,* Vol. 36, pp. 1175–1186, September 1989.

R. T. Baird and T. S. Fiez. "Improved Delta-Sigma DAC Linearity Using Data-Weighted Averaging," *Proc. IEEE Symp. on Circuits and Systems,* pp. 13–16, Seattle, 1995.

B. E. Boser and B. A. Wooley. "The Design of Sigma-Delta Modulation Analog-To-Digital Converters," *IEEE J. of Solid-State Circuits,* Vol. 23, pp. 1298–1308, December 1988.

J. C. Candy. "Decimation for Sigma-Delta Modulation," *IEEE Trans. Commun.,* Vol. 34, pp. 72–76, January 1986.

J. C. Candy and G. C. Temes. "Oversampling Methods for A/D and D/A Conversion," *Oversampling Delta-Sigma Data Converters,* ed. J. C. Candy and G. C. Temes, IEEE Press, New York, 1992.

L. R. Carley. "A Noise-Shaping Coder Topology for 15+ Bit Converters," *IEEE J. of Solid-State Circuits,* Vol. 24, pp. 267–273, April 1989.

K. C.-H. Chao, S. Nadeem, W. L. Lee, and C. G. Sodini. "A Higher-Order Topology for Interpolative Modulators for Oversampling A/D Converters," *IEEE Trans. on Circuits and Systems,* Vol. 37, pp. 309–318, March 1990.

F. Chen and B. H. Leung. "A High-Resolution Multibit Sigma-Delta Modulator with Individual Level Averaging," *IEEE J. of Solid-State Circuits,* Vol. 30, pp. 453–460, April 1995.

J. Dattorro et al. "The Implementation of One-Stage Multirate 64:1 FIR Decimator for Use in One-Bit Sigma-Delta A/D Applications," *AES 7th Int. Conf.,* May 1989.

P. Ferguson, Jr., et al. "An 18b 20kHz Dual Sigma-Delta A/D Converter," *Int. Solid-State Circuits Conf.,* February 1991.

I. Galton. "Spectral shaping of circuit errors in digital-to-analog converters," *IEEE Trans. on Circuits and Systems—II: Analog and Digital Signal Processing,* vol.44, pp. 808–817, Oct. 1997.

A. Hairapetian and G. C. Temes. "A Dual-Quantization Multi-Bit Sigma-Delta Analog/Digital Converter," *IEEE Int. Symp. on Circuits and Systems,* pp. 437–441, May 1994.

A. Hairapetian, G. C. Temes, and Z. X. Zhang. "Multibit Sigma-Delta Modulator with Reduced Sensitivity to DAC Nonlinearity," *Electron. Lett.,* Vol. 27, pp. 990–991, May 1991.

S. A. Jantzi, W. M. Snelgrove, and P. F. Ferguson, Jr. "A Fourth-Order Bandpass Sigma-Delta Modulator," *IEEE J. of Solid-State Circuits,* Vol. 28, pp. 282–291, March 1993.

P. Kiss, J. Silva, A. Wiesbauer, T. Sun, U.-K. Moon, J. T. Stonick, and G. C. Temes, "Adaptive digital correction of analog errors in MASH ADCs—Part II: Correction using test-signal injection," *IEEE Trans. Circuits Syst. II*, pp. 629–638, July 2000.

L. E. Larson, T. Cataltepe, and G. C. Temes. "Multibit Oversampled $\Sigma - \Delta$ A/D Converter with Digital Error Correction," *Electron. Lett.,* Vol. 24, pp. 1051–1052, August 1988.

H. A. Leopold et al. "A Monolithic CMOS 20-b Analog-to-Digital Converter," *IEEE J. of Solid-State Circuits,* Vol. 26, pp. 910–916, July 1991.

T. C. Leslie and B. Singh. "An Improved Sigma-Delta Modulator Architecture," *IEEE Int. Symp. on Circuits and Systems,* pp. 372–375, May 1990.

B. Leung, "Multibit sigma-delta A/D converters incorporating a novel class of dynamic element matching techniques," *IEEE Trans. on Circuits and Systems—II: Analog and Digital Signal Processing,* vol. 39, pp. 35–51, Jan. 1992.

Y. Matsuya, et al. "A 16-bit Oversampling A-to-D Conversion Technology Using Triple-Integration Noise Shaping," *IEEE J. of Solid-State Circuits,* Vol. 22, pp. 921–929, December 1987.

S. Norsworthy. "Optimal Nonrecursive Noise-Shaping Filters for Oversampling Data Converters–Part I and II," *IEEE Int. Symp. on Circuits and Systems,* pp. 1353–1360, Chicago, May 1993.

T. Saramaki et al. "Multiplier-Free Decimator Algorithms for Super-Resolution Oversampled Converters," *IEEE In t. S ymp. on Circuits and Systems,* pp. 3275–3278, New Orleans, May 1990.

M. Sarhang-Nejad and G. C. Temes, "A High-Resolution Multibit $\Sigma\Delta$ ADC with Digital Correction and Relaxed Amplifier Requirements," *IEEE J. of Solid-State Circuits,* Vol. 28, pp. 648–660, June 1993.

H. J. Schouwenaars, D. W. J. Groeneveld, C. A. A. Bastiaansen, and H. A. H. Termeer. "An Oversampled Multibit CMOS D/A Converter for Digital Audio with 115-dB Dynamic Range," *IEEE J. of Solid-S tate Circuits,* Vol. 26, pp. 1775–1780, December 1991.

R. Schreier. "An Empirical Study of High-Order Single-Bit Delta-Sigma Modulators," *IEEE Trans. on Circuits and Systems—II: Analog and Digital Signal Processing,* Vol. 40, pp. 461–466, August 1993.

R. Schreier and W. M. Snelgrove. "Bandpass Sigma-Delta Modulation," *Electron. Le tt.,* Vol. 25, no. 23, pp. 1560–1561, November 9, 1989.

G. C. Temes, S. Shu, and R. Schreier. "Architectures for $\Delta\Sigma$ DACs," in *Oversampling $\Delta\Sigma$ Data Converters,* ed. G. C. Temes, IEEE Press, 1996.

P. P. Vaidyanathan. *Multirate Systems and Filte r Banks.* Prentice Hall, Englewood Cliffs, New Jersey, 1993.

18.12 PROBLEMS

18.12.1 Section 18.1: Oversampling without Noise Shaping

18.1 A converter with no noise shaping exhibits a SNR of 40 dB over a 1-MHz signal bandwidth. By what factor must the sampling rate be increased (without noise shaping) to obtain a maximum SNR of 80 dB?

18.2 A white noise source has a power of 1 mW when sampled at $f_s = 10$ MHz. If the same signal is sampled at $f_s = 100$ MHz and the result low-pass filtered with a bandwidth of $f_s/20$, what is the resulting noise power?

18.3 Given that an 8-bit A/D converter has a SNR of 50 dB but is linear to 12 bits, what is the sampling rate required to achieve 12 bits of accuracy using straight oversampling on a signal bandwidth of 1 MHz?

18.12.2 Section 18.2: Oversampling with Noise Shaping

18.4 Repeat Example 18.4, except use a dc input level of $1.2/3$. At what frequency would a tone appear (relative to the sampling rate, f_s)?

18.5 Repeat Example 18.4, except use a dc input level of 1.1 to see if the internal state of the modulator saturates.

18.6 Repeat Example 18.4, except use an input sequence of $\{10, -10, 10, -10, 10, -10, \ldots\}$ to see if the internal state of the modulator saturates. This problem demonstrates that a large level input signal can be applied to a modulator if its signal power resides at a frequency where the gain of $H(z)$ is low.

18.7 Repeat Problem 18.3, assuming that the 8-bit A/D converter is placed inside a first-order delta-sigma modulator. What is the sampling rate if a second-order modulator is used?

18.8 Assuming a 1-bit modulator has an output signal of ± 1 and the input sinusoidal signal is ± 0.5 peak-to-peak, what is the ratio of the signal power to quantization noise power before any filtering is applied?

18.9 What does the block $G(z) - 1$ need to be in Fig. 18.13 in order that a second-order noise transfer function is realized equal to $(1 - z^{-1})^2$?

18.10 Assume that the switched capacitor realization of the first-order A/D modulator in Fig. 18.9 introduces thermal noise at a rms input-referred level of $\sqrt{2kT/C}$. Using a one-bit quantizer, and allowing equal noise contributions from the thermal noise and quantization noise, find the required value of C (at a temperature 300 K) and of OSR in order to achieve an SNR of 90 dB.

18.11 Show a switched-capacitor modulator realizing the second-order modulator shown in Fig. 18.10.

18.12.3 Section 18.3: System Architectures

18.12 The analog input x_{in} to the oversampled A/D converter system shown in Fig. 18.14 contains a signal power of $0.25\ V^2$ within a signal bandwidth of $f_0 = 1$ MHz and white noise with a spectral density of $50\ nV/\sqrt{Hz}$. Assuming a first-order anti-aliasing filter is used, what is the SNR at the input to the modulator, x_{sh}? Assuming a 1-bit first-order modulator, what OSR is required to ensure that quantization noise is less than the input noise?

18.13 An oversampled D/A converter utilizes second order noise shaping and a sampling rate of $f_s = 200$ MHz with $OSR = 256$. What should be the bandwidth and order of the following analog smoothing filter? Assuming a Butterworth smoothing filter is used, estimate the attenuation at $f_s/2$.

18.12.4 Section 18.4: Digital Decimation Filters

18.14 Show that the sequence of $\{1, 1, -1, 1, 1, -1, 1, 1, -1, \ldots\}$ into a cascade of three integrators and differentiators shown in Fig. 18.19 gives the correct running average. Assume that the adders have infinite width and that $M = 4$.

18.15 At first glance, the impulse response of a running average filter of length 4 might appear the same as that for a hold system that holds the input signal for length 4. By producing a block diagram for each system, explain the difference between the two systems.

18.12.5 Section 18.5: Higher-Order Modulators

18.16 You are to realize a third-order MASH by cascading three first-order stages as shown in Fig. P18.16. The second and third stages accept as inputs the quantization errors of the previous stages, $e_{1,2}(n)$. Assuming all of the discrete-time integrators are ideal, find the digital filters H_1, H_2, and H_3 that should be applied to the outputs of each of the three modulators before summing them so that the overall MASH output y includes only the input signal u and third-order shaped quantization noise.

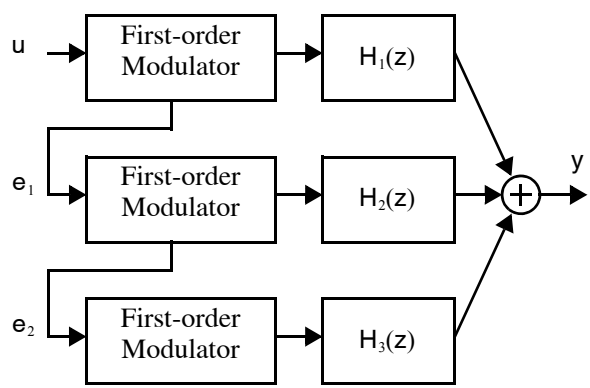


Fig. P18.16

18.17 Show the block diagram for a MASH architecture that realizes third-order noise shaping using a second-order modulator (as shown in Fig. 18.10) and a first-order modulator.

18.12.6 Section 18.6: Bandpass Oversampling Converters

18.18 Sketch the noise transfer function from 0 to 4 MHz for the second-order bandpass modulator shown in Fig. 18.23, assuming the sampling rate, f_s, equals 4 MHz.

18.19 Show how the block diagram in Fig. 18.23 can be modified so that the bandpass modulator places zeros at $\pm f_s/8$ rather than $\pm f_s/4$.

18.12.7 Section 18.7: Practical Considerations

18.20 Using the "typical" waveform pattern shown in Fig. 18.24, find the practical average outputs when 0, $1/2$, and $-1/2$ have been applied to the first-order modulator shown in Fig. 18.7.

18.21 A single-bit second-order oversampled A/D converter is to be realized with a SQNR of 80 dB. Estimate a lower bound on the required dc gain of the opamps that can be used in its implementation.

18.12.8 Section 18.8: Multi-Bit Oversampling Converters

18.22 We wish to model a nonlinear 2-bit D/A, with output levels given by {–0.61, –0.17, 0.25, 0.59} as an ideal D/A followed by an additive error, Q_d. Assuming that the four levels occur with equal probability, what is the power of the ideal D/A quantization noise, and what is the error power due to the nonlinearity, Q_d?

18.23 Analyze the third-order modulator in Fig. 18.29, proving (18.55) and (18.56).

CHAPTER

19 Phase-Locked Loops

Practically all digital and radio frequency circuits and most analog circuits require a precision oscillator. Unfortunately, integrated circuit oscillators are not, on their own, suitable for use as frequency- or time-references in high-performance circuits. A major problem is that their frequency of oscillation is not known precisely. Furthermore, the jitter of integrated circuit oscillators (which can be thought of as random fluctuations in frequency) is too high for many applications. Therefore, integrated circuit oscillators are most often inserted into a phase locked loop where their operating frequency and phase is continuously compared against a precise external time-reference and automatically tuned to maintain a constant frequency and phase. The process is analogous to generating a precise on-chip voltage or current based on an externally-supplied reference. In summary, a phase-locked loop is a circuit that uses an external timing reference to measure and generate on-chip frequencies and times accurately.

It is natural to wonder why, if a precise time-reference is available, a phase locked loop is required at all. Why not simply use the time-reference (usually itself a clock) directly? One reason is that the required frequency may be different than that of the reference. This is frequently the case since timing references are commonly derived from precision resonators trimmed to a fixed frequency. For low cost, these resonators are often in the range of 1 – 200 MHz. The generation of a different and/or variable-frequency precision clock from a fixed-frequency reference is referred to as *clock synthesis* or *frequency synthesis* and is a major application of phase-locked loops.

In some cases, the timing reference is supplied by another integrated circuit with which the circuit under design must communicate. Frequency variations can be caused by changes in temperature, supply voltage or aging of components, or can be introduced intentionally to reduce radiated emissions. Phase-locked loops are required to ensure the two circuits track each other very accurately, both in frequency and phase, so that they may communicate with each other reliably. Finally, phase-locked loops are sometimes used to frequency-tune circuits such as filters.

19.1 BASIC PHASE-LOCKED LOOP ARCHITECTURE

The basic architecture of a phase-locked loop (PLL) is shown in Fig. 19.1. It usually consists of a phase detector, a low-pass filter called the loop filter, a voltage controlled oscillator (VCO), and a clock divider circuit configured in a loop. The phase detector is a circuit that normally has an output voltage or current with an average value proportional to the phase difference between the input (reference) signal and the output of the VCO. The loop filter is used to extract the average value from the output of the phase detector. This average value is then used to drive the VCO, whose output frequency changes in response to changes in V_{cntl}. Since the reference input is usually at a lower frequency than the VCO output clock, a clock divider circuit is used to reduce the output clock frequency prior to comparison by the phase detector. The negative feedback of the loop results in the input signal being synchronized with the output of the divider. This basic architecture is called an *integer-N PLL* since, in a stable steady-state, the divider output and input reference will have the same frequency implying that the VCO output frequency must be exactly an integer multiple N of the input frequency.

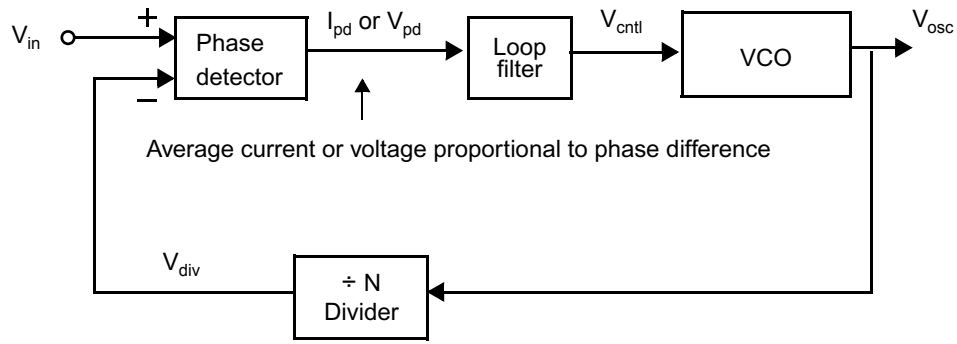


Fig. 19.1 The basic architecture of a phase-locked loop.

A PLL is a negative feedback system, similar to a negative feedback amplifier. In a negative feedback amplifier, two voltages or currents are subtracted and the loop's large gain ensures that their difference is made small. In a phase-locked loop, the objective is to make the difference between the phase of the VCO output and a reference input clock very small. Hence, at the PLL input these two quantities are subtracted by the phase detector and the result is applied to a large gain through the loop filter and VCO to ensure that in a stable steady-state the difference is small.

19.1.1 Voltage-Controlled Oscillator

A key component of PLLs is the VCO, a circuit with an oscillatory output (i.e., sinusoid or other clock signal) whose frequency depends upon an input control voltage, V_{cntl}. Hence, assuming the VCO output is a sinusoidal voltage slowly varying with time, it can be represented by the equation

$$V_{osc}(t) = E \sin[\omega_0 t + \phi(t)] \tag{19.1}$$

where ω_0 is the oscillator's nominal or *free-running* frequency, and the oscillator's instantaneous phase is $\phi(t)$. The oscillator's instantaneous frequency is the rate of change of $[\omega_0 t + \phi(t)]$ in units of rad/sec.

$$\omega_{inst}(t) = \frac{d[\omega_0 t + \phi(t)]}{dt} = \omega_0 + \frac{d\phi(t)}{dt} \tag{19.2}$$

Hence, if the oscillator phase $\phi(t)$ is constant, the instantaneous frequency is equal to the free-running frequency, $\omega_{inst}(t) = \omega_0$.

Now define the deviation in the VCO's instantaneous frequency from its free running frequency,

$$\omega(t) = \omega_{inst}(t) - \omega_0 \tag{19.3}$$

Of course, $\omega(t)$ is dependent on the VCO control voltage $V_{cntl}(t)$. Clearly from (19.2), $\omega(t)$ is linearly related to the phase.

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{19.4}$$

If we integrate both sides of (19.4), we have

$$\phi(t) = \phi(0) + \int_0^t \omega(\tau)\, d\tau \tag{19.5}$$

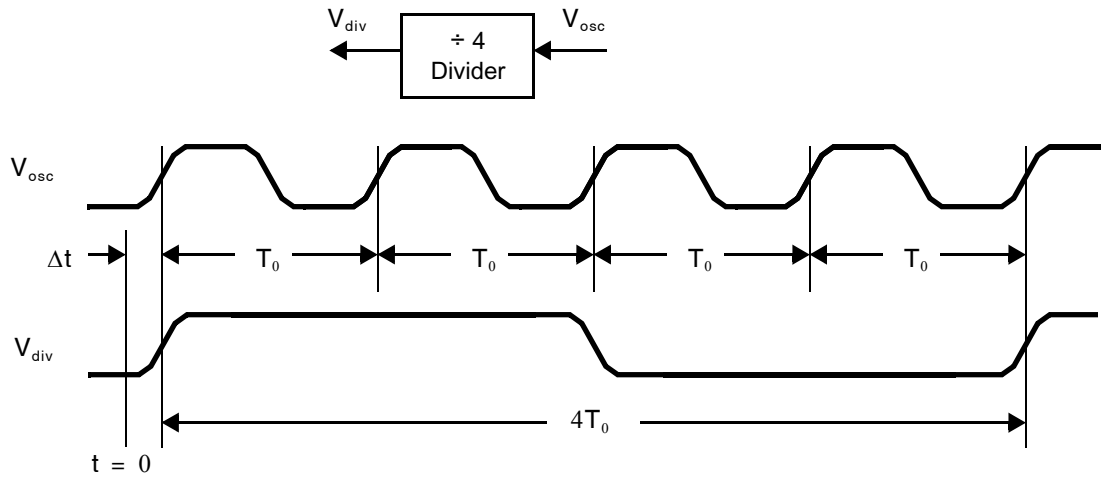


Fig. 19.2 The operation of a clock divider with a division ratio of $N = 4$.

Key Point: *The instantaneous frequency deviation of a VCO from its free-running frequency is the derivative of its instantaneous phase. Hence, its instantaneous phase is the integral of the instantaneous frequency deviation.*

and we see that the *instantaneous phase is the integral of the instantaneous frequency deviation*. Taking the Laplace transform of both sides of (19.4), we have

$$\phi(s) = \frac{\omega(s)}{s} \tag{19.6}$$

where $\phi(s)$ and $\omega(s)$ are the Laplace transforms of $\phi(t)$ and $\omega(t)$ respectively. Equation (19.6) will be used to model the relationship between the VCO's input control voltage and the phase of its output signal. To arrive at a linear model, assume that $\omega(t)$ is proportional to the control voltage with a constant of proportionality K_{osc} that has units $\text{rad}/(\text{sec} \cdot \text{V})$.

$$\omega(t) = K_{osc} V_{cntl}(t) \tag{19.7}$$

Even when the relationship is nonlinear, (19.7) can still be used as a small-signal model where K_{osc} relates small changes in $V_{cntl}(t)$ to the resulting changes in $\omega(t)$.

19.1.2 Divider

An ideal clock divider with a division ratio N will produce a cyclic output with each period corresponding to exactly N periods of the input. This is illustrated in Fig. 19.2 for $N = 4$. The result is that the divider's instantaneous output frequency is exactly $(1/N)$ times that of is input, which in PLLs is the VCO output. Furthermore, the divider's output phase, $\phi_{div}(t)$, is also exactly $(1/N)$ times that of is input, $\phi(t)$. To see this, note that the phase of the VCO output V_{osc} is the time delay Δt between that clock's rising edge and some reference time, $t = 0$, normalized with respect to a period, T_0.

$$\phi = 2\pi \frac{\Delta t}{T_0} \tag{19.8}$$

Assuming the divider output is synchronized with that of the VCO output, its phase is determined by the same time delay, Δt, but normalized with respect to the longer period, NT_0.

$$\phi_{div} = 2\pi \frac{\Delta t}{NT_0} = \frac{\phi}{N} \tag{19.9}$$

Equation (19.9) provides a linear model for the divider, which in terms of Laplace transforms is simply

$$\phi_{div}(s) = \frac{\phi(s)}{N} \tag{19.10}$$

19.1.3 Phase Detector

The role of the phase detector is to produce a signal whose average value is proportional to the difference between the phase of the reference input, θ_{in}, and the feedback clock signal, θ_{div}. This is analogous to the subtraction performed at the input of a feedback amplifier. There are many possibilities for the phase detector: a simple exclusive-OR gate, a sample and hold, an analog multiplier, or a combination of digital circuits such as D flip-flops. All accept two clock signals as inputs and produce an output whose average value provides the requisite phase information.

EXAMPLE 19.1

What is the gain of an analog multiplier phase detector?

Solution

In this example, the input signal, V_{in}, is assumed to be a sinusoid with a known amplitude. The phase detector is realized as an analog multiplier having a relationship given by

$$V_{pd} = K_M V_{in} V_{div} \tag{19.11}$$

where V_{div} is the divider output, V_{pd} is the phase-detector output, and K_M is a multiplication constant. Let the input signal be described by a sinusoid having

$$V_{in} = E_{in} \sin(\omega t) \tag{19.12}$$

and the output of the divider is given by a sinusoid where

$$V_{div} = E_{div} \sin(\omega t - \phi_d + 90°) = E_{div} \cos(\omega t - \phi_d) \tag{19.13}$$

The argument ϕ_d represents the phase difference between the input signal and the output of the oscillator. The reason for the $90°$ offset is so that when $\phi_d = 0$ the average output of the phase detector (an analog multiplier in this example) is zero. The combination of an analog multiplier followed by a low-pass filter performs what is commonly known as the correlation function on analog signals. When the input signal and the VCO output have a $90°$ phase difference (i.e., $\phi_d = 0$), they are uncorrelated and the output of the low-pass filter will be zero. From this $90°$ phase offset, if the phase difference changes so that the two waveforms are more in phase (i.e., $\phi_d > 0$), then the correlation and, therefore, the output of the low-pass filter will be positive. For example, from (19.13), if $\phi_d = 90°$, then the output of the oscillator and the input signal would be exactly in phase, and the average output of the phase detector is found to be the amplitudes of the two sinusoids multiplied together and divided by two, resulting in

$$V_{cntl} = K_{lp} K_M \frac{E_{in} E_{div}}{2} \tag{19.14}$$

This situation is approximately shown in Fig. 19.3 where ϕ_d is taken slightly smaller than $90°$ to distinguish the two waveforms. Also shown is the output of the multiplier, V_{pd}, for the case of $E_{in} = E_{div} = 1$ V, and

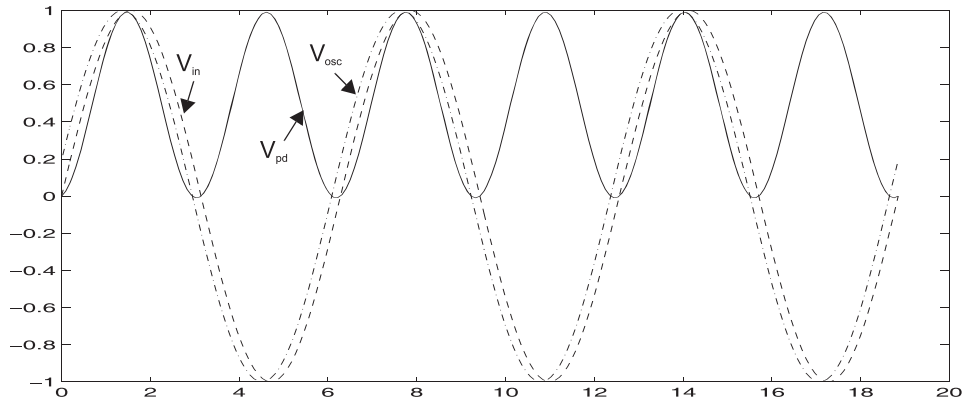


Fig. 19.3 The output of the phase detector in Example 19.1 when the input signal and the oscillator are nearly in phase.

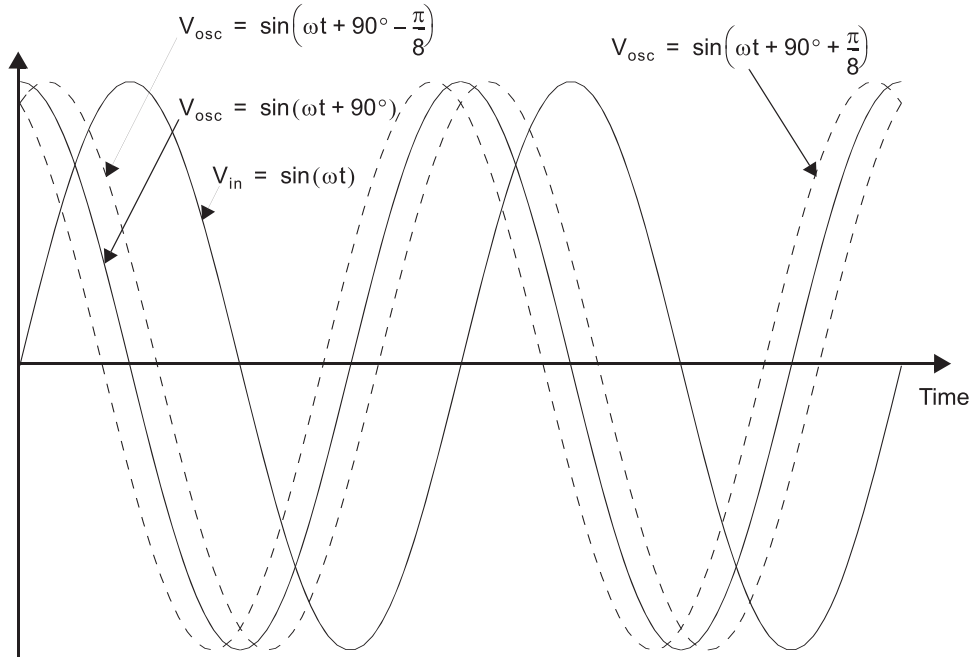


Fig. 19.4 The relative phase angles of the input signal and the output of the oscillator.

$K_M = 1\ V^{-1}$. Here, it is seen that the output is at twice the frequency of the input signal and that it has an offset voltage of approximately $1/2$ V. This corresponds closely to the maximum possible output of the low-pass filter. On the other hand, if the phase changes so as to make the two waveforms more out of phase, then the correlation between them and the output of the low-pass filter will be negative. Shown in Fig. 19.4 are some examples of the different waveforms. It is readily seen that $\phi_d > 0$ corresponds to the waveforms that are more in phase.

These statements can be quantified analytically. The output of the phase detector is given by

$$V_{pd} = K_M V_{in} V_{div} = K_M E_{in} E_{div} \sin(\omega t) \cos(\omega t - \phi_d) \tag{19.15}$$

Using the trigonometric identity

$$\sin(A) \cos(B) = \frac{1}{2}[\sin(A + B) + \sin(A - B)] \tag{19.16}$$

we have

$$V_{pd} = K_M \frac{E_{in} E_{div}}{2}[\sin(\phi_d) + \sin(2\omega t - \phi_d)] \tag{19.17}$$

The low-pass filter removes the second term at twice the frequency of the input signal.[1] The signal V_{cntl} is therefore given by

$$V_{cntl} = K_{lp} K_M \frac{E_{in} E_{div}}{2} \sin(\phi_d) \tag{19.18}$$

Note that V_{cntl} is either a dc value or a slowly varying signal, assuming that the frequency of the input signal is constant or slowly varying. For small ϕ_d, we can approximate (19.18) as

$$V_{cntl} \cong K_{lp} K_M \frac{E_{in} E_{div}}{2} \phi_d = K_{lp} K_{pd} \phi_d \tag{19.19}$$

Thus, the output of the low-pass filter is approximately proportional to the phase difference between the output of the oscillator and the input signal, assuming the $90°$ offset bias is ignored. This approximation is used in analyzing the PLL to obtain a linear model. The constant of proportionality is called K_{pd} and is given by

$$K_{pd} = K_M \frac{E_{in} E_{div}}{2} \tag{19.20}$$

Note that with this analog-multiplier phase detector, K_{pd} is a function of the input amplitude and divider output amplitude. As we shall soon see, K_{pd} affects loop dynamics, and thus we see the need for having known values for E_{in} and E_{div}.

The output of the phase detector may be a voltage as in Example 19.1, but in CMOS circuits is commonly a current, either single-ended or differential. When the output is a current whose average value[2] is I_{pd}, the gain of the phase detector is a constant K_{pd} with units A/rad.

$$I_{pd} = K_{pd} \phi_d = K_{pd}(\phi_{in} - \phi_{div}) \tag{19.21}$$

EXAMPLE 19.2

A charge-pump phase detector is shown in general form in Fig. 19.5. The phase detector logic is designed so that logic-high pulses are generated on P_u when the reference clock leads the feedback clock, $\phi_d > 0$, and pulses are generated on P_d when the reference clock is lagging, $\phi_d < 0$. A typical arrangement of logic is shown in Fig. 19.6. Find the gain of such a phase detector K_{pd} in units A/rad.

---

1. The extent to which this is not true is a major source of distortion when sinusoidal VCOs are used.
2. This "average" is recomputed each period of the reference clock since, of course, the value of I_{pd} will vary over time.

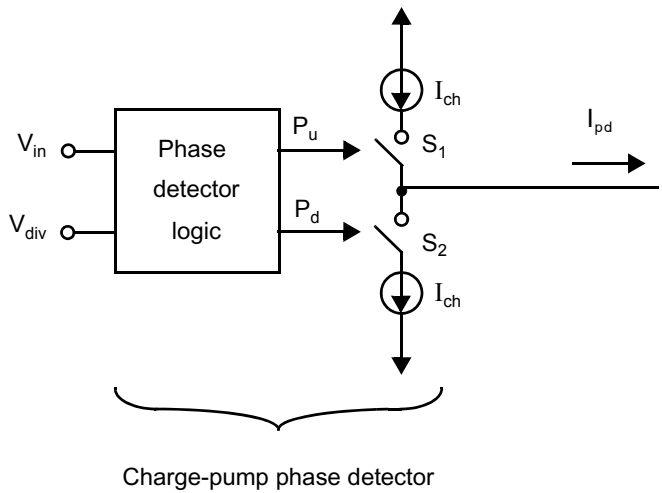


Fig. 19.5 General charge-pump phase detector.

Solution

This phase comparator either sources or sinks current in the low-pass loop filter, depending on the output of the sequential phase-detector logic. When S_1 is closed, I_{ch} flows into the low-pass filter; when S_2 is closed, I_{ch} flows out of the low-pass filter. The sequential phase-detector logic generates two outputs, P_u and P_d, depending on its inputs, V_{in} and V_{div}. These are normally digital signals from the PLL input and the VCO, respectively.[3] Although there are many possibilities for the sequential phase detector, the one we will consider is controlled by the leading edges of V_{in} and V_{osc}. If V_{in} goes high before V_{osc}, then P_u will be high during the time that the signals are different. This will inject some charge into the loop filter to speed up the VCO. Similarly, if V_{osc} goes low before V_{in}, then P_d will be high during the time the inputs are different, S_2 will be closed, and charge will be subtracted from the loop filter, which decreases the VCO frequency. When the leading edges are coincident, the two inputs are phase aligned. In this case, ideally both P_u and P_d are always low, both S_1 and S_2 are always open so no charge is injected to or removed from the loop filter, and the loop has reached steady-state. In practical sequential phase detectors, both P_u and P_d may go high for a few gate delays when the input waveforms are coincident, but this time is usually short enough that it has very little effect on the dc output of the loop filter, and its effect is largely eliminated by capacitors in the loop filter. The falling edges of both V_{in} and V_{osc} have no effect on P_u and P_d, which are both zero at this time. This makes the PLL insensitive to the duty cycles of the waveform. Typical waveforms are shown in Fig. 19.6.

To solve for the phase detector gain, note that the output current I_{pd} averaged over any single period of the reference clock V_{in} is given by the charge pump current I_{ch} scaled by the fraction of a period over which the phase detector outputs P_u and P_d are active.

$$\overline{I_{pd}} = I_{ch}\frac{\phi_d}{2\pi} \tag{19.22}$$

3. This PLL is only used when the VCO output and the reference input are both digital signals or have been converted to digital signals by a limiter.

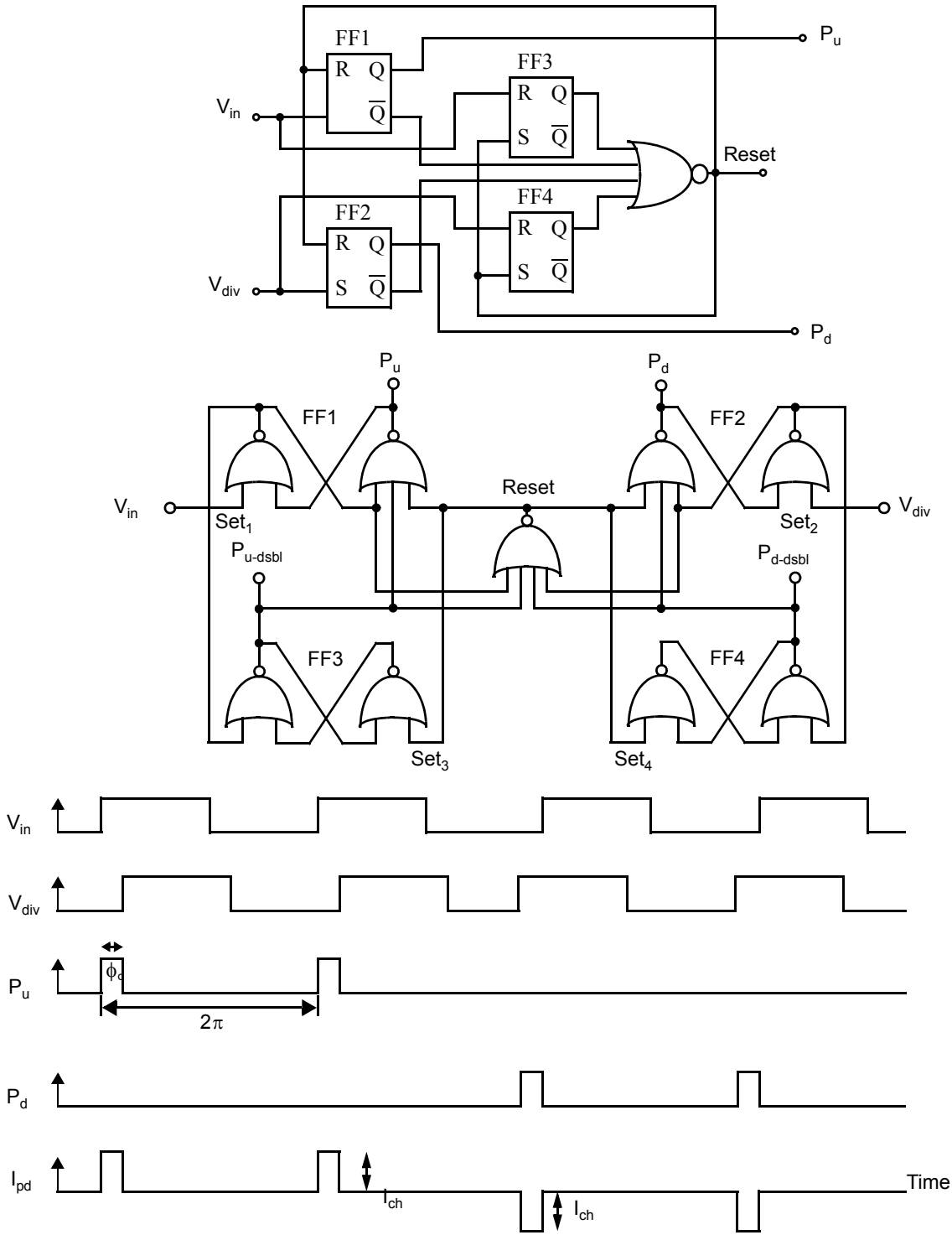


Fig. 19.6 Phase-detector logic and corresponding waveforms.

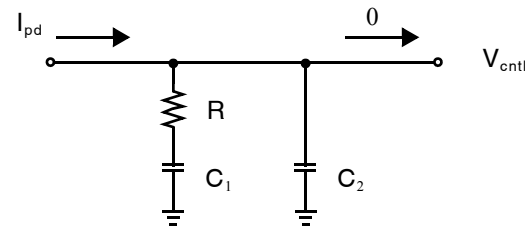


Fig. 19.7 Loop filter for charge-pump phase detectors.

Combining (19.21) and (19.22) yields the phase detector gain,

$$K_{pd} = \frac{I_{ch}}{2\pi} \tag{19.23}$$

19.1.4 Loop Filer

The loop filter plays several key roles in a PLL. It averages out the quickly-varying components of the phase detector output. We shall see that it also influences the dynamics of the loop by determining how the VCO control signal, V_{cntl}, responds to variations in the phase difference represented by I_{pd}.

In most PLLs, the loop filter is first or second order. For example, a common first-order low-pass loop filter has a transfer function given by

$$K_{lp}H_{lp}(s) = K_{lp}\frac{1 + s/\omega_z}{s} = K_{lp}\left(\frac{1}{s} + \frac{1}{\omega_z}\right) \tag{19.24}$$

A loop filter of the form (19.24) may be thought of as comprising the sum of a proportional and an integral path. Many loop filters will include additional higher-order poles and zeros, but the most important PLL design parameters are captured by (19.24).

The loop filter gain constant K_{lp} has units of $\Omega \cdot \text{rad}/s = F^{-1}$ when the phase detector output is a current and rad/s when the phase detector output is a voltage.

EXAMPLE 19.3

Find the loop filter transfer function that relates the current I_{pd} and control voltage V_{cntl} in Fig. 19.7.

Solution

It is assumed that there is no current flowing out of the loop filter into the control terminal of the VCO, as shown in Fig. 19.7. Hence, all current must flow through the parallel impedances $(R + 1/(sC_1))$ and $1/(sC_2)$. The loop filter transfer function is

$$K_{lp}H_{lp}(s) = \frac{V_{cntl}(s)}{I_{pd}(s)} = \frac{1}{s(C_1 + C_2)} \cdot \frac{1 + sRC_1}{1 + sR\left(\dfrac{C_1C_2}{C_1 + C_2}\right)} \tag{19.25}$$

In practice, the capacitor C_2 is there only to suppress voltage spikes that would otherwise arise across R from transient currents, hence its value is generally much smaller than C_1. Assuming that $(C_1 + C_2) \cong C_1$ and neglecting the high frequency pole at $(C_1 + C_2)/(RC_1C_2)$,

$$K_{lp}H_{lp}(s) = \frac{V_{cntl}(s)}{I_{pd}(s)} \cong \frac{1}{C_1}\left(\frac{1 + sRC_1}{s}\right) \tag{19.26}$$

Casting (19.26) into the general model of (19.24), we have for a charge-pump loop filter,

$$K_{lp} = 1/C_1 \tag{19.27}$$

$$\omega_z = 1/(RC_1) \tag{19.28}$$

The loop filter in Fig. 19.7 is often combined with the charge pump phase comparator in Fig. 19.5. When S_1 is closed, I_{ch} flows into the low-pass filter, increasing the control voltage into the VCO; when S_2 is closed, I_{ch} flows out of the low-pass filter, decreasing the control voltage of the VCO. When both switches are open, the top plates of C_1 and C_2 are open circuited and the output voltage remains constant in the steady state. The resistor, R, has been included to realize a zero in the low-pass filter's transfer function. It provides an output component proportional to the input current, while the capacitor integrates the input current. Since the two are in series, the sum of the proportional and integral terms are seen at V_{cntl}.

19.1.5 The PLL in Lock

To understand phase-locked loops, a simple example will be considered. Assume a division ratio $N = 1$ so that $\phi_{div} = \phi$. Furthermore, assume an input signal frequency initially equal to the free-running frequency of the VCO, the system is initially in lock with $\phi_d = 0$, and the output of the low-pass filter V_{cntl} is also equal to zero. Next, assume the input frequency slowly increases. This will cause it to lead the VCO output, which corresponds to a phase difference $\phi_d = \phi_{in} - \phi > 0$. After a short time (roughly the time constant of the low-pass filter) the output of the low-pass filter will go positive. Since the two waveforms are now assumed to be at slightly different frequencies, the phase difference, and therefore the output of the low-pass filter, will slowly continue to increase. However, the frequency of the VCO is proportional to its control voltage, so the increase in the low-pass filter's output will cause the VCO's frequency to increase until it is the same as that of the input signal again, which will keep the two signals in synchrony (i.e., locked). Of course, the opposite would occur if the input signal's frequency decreased. Specifically, the average output of the phase detector would become negative, which, after being averaged by the low-pass filter, would drive the oscillator to a lower frequency until it is again identical to the frequency of the input signal and the two waveforms are again synchronized. It is readily seen here that the phase-locked loop stays in lock because of the negative feedback of the loop in which the phase of the oscillator's output is subtracted from the phase of the input signal.

The argument is easily extended to the case $N \neq 1$. The divider ensures that the oscillator output is at a phase and frequency N-times greater than the clock that is fed back to the phase detector. Negative feedback and high dc gain around the loop force the two phases at the inputs of the phase detector to be equal. Hence, it is seen that a PLL in lock produces an output clock at a frequency precisely N-times greater than the reference input that tracks phase changes in the reference input.

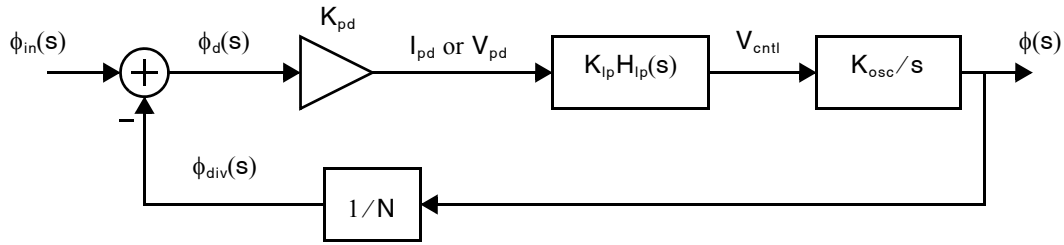


Fig. 19.8 A signal-flow graph for the linearized small-signal model of a PLL when in lock.

19.2 LINEARIZED SMALL-SIGNAL ANALYSIS

Key Point: *Once a PLL is in lock, its dynamic response to input phase and frequency changes can be well approximated by a linear model, as long as these changes are slow and small about their operating or bias point.*

The dynamics of a PLL will be of great interest to the designer. Specifically, is the loop stable? How does it settle when perturbed from its locked point? How is noise generated within the loop translated to its output? How can the loop be designed to ensure a certain dynamic behavior? As with feedback amplifiers, linear analysis is a useful tool to answer these questions. However, the phase-locked loop is a highly nonlinear system. Fortunately, once a PLL is in lock, its dynamic response to input phase and frequency changes can be well approximated by a linear model, as long as these changes are slow and small about their operating or bias point.

Using the relationships described in Section 19.1, a linear model for a phase-locked loop can be developed. It should be noted that this is a small-signal model in that it relates changes about an operating point. In other words, the phase variable represents *changes in phase* from reference signals that were identical to the true signals at the beginning of the analysis. Similarly, any frequency variables, which will be introduced shortly, also represent *changes from the original frequencies* at the beginning of the analysis. A signal-flow graph for the linearized small-signal model is shown in Fig. 19.8.

It is now possible to analyze the signal flow graph of Fig. 19.8 to determine transfer functions that relate small perturbations in the various signals around the loop. For example, straightforward feedback analysis reveals that the transfer function relating small changes in the input phase $\phi(s)$ to the phase difference $\phi_d(s)$ is given by

$$\frac{\phi_d(s)}{\phi_{in}(s)} = \frac{1}{1 + L(s)} \tag{19.29}$$

where $L(s)$ is the loop gain,

$$L(s) = \frac{K_{pd}K_{lp}K_{osc}H_{lp}(s)}{Ns} \tag{19.30}$$

Substituting (19.30) into (19.29),

$$\frac{\phi_d(s)}{\phi_{in}(s)} = \frac{s}{s + K_{pd}K_{lp}K_{osc}H_{lp}(s)/N} \tag{19.31}$$

Equation (19.31) is general in that it applies to almost every phase-locked loop. Differences between one PLL and another are determined only by what is used for the low-pass filter (which determines $H_{lp}(s)$), the phase detector (which determines K_{pd}), or the oscillator (which determines K_{osc}).

19.2.1 Second-Order PLL Model

Assume the loop filter has the transfer function given in (19.24). Define the loop constant in units rad/s,

$$\omega_{pll} = \sqrt{\frac{K_{pd}K_{lp}K_{osc}}{N}} \tag{19.32}$$

Substituting (19.24) and (19.32) into (19.30) and (19.31) and rearranging results in

$$L(s) = \frac{\omega_{pll}^2}{s^2} \cdot \left(1 + \frac{s}{\omega_z}\right) \tag{19.33}$$

$$\frac{\phi_d(s)}{\phi_{in}(s)} = \frac{1}{\omega_{pll}^2} \cdot \frac{s^2}{\left(1 + \frac{s}{\omega_z} + \frac{s^2}{\omega_{pll}^2}\right)} \tag{19.34}$$

The dc gain of (19.34) is zero which signifies that the phase difference between the input and divider output will eventually settle to zero.

We can also relate small changes in the phase of the input and output clock.

$$H(s) = \frac{\phi(s)}{\phi_{in}(s)} = \frac{N(1 + s/\omega_z)}{1 + \frac{s}{\omega_z} + \frac{s^2}{\omega_{pll}^2}} \tag{19.35}$$

The transfer function $H(s)$ may be referred to as the phase transfer function or *jitter transfer function* of the PLL. The dc gain of $H(s)$ is N. This implies that when the input phase changes by a small amount, $\Delta\phi_{in}$, the output phase changes (after an initial transient) by $N\Delta\phi$. Since the output frequency is N-times larger, this means that the input and output clocks are tracking one another.

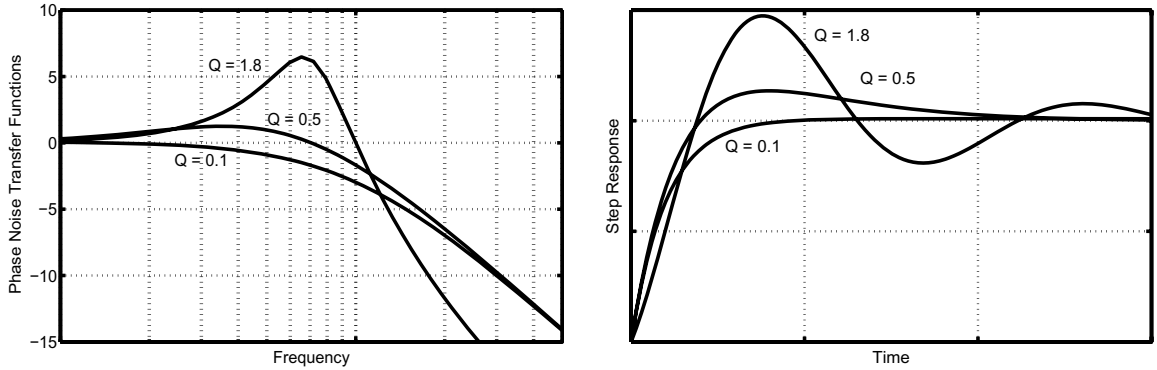
In the case of a first-order loop filter both (19.34) and (19.35) are second-order transfer functions and the PLL is a second-order system. Equating the denominator with the general form $1 + s/(\omega_r Q) + s^2/\omega_r^2$ reveals that the loop has a resonant frequency $\omega_r = \omega_{pll}$ and a Q-factor[4] given by

$$Q = \frac{\omega_z}{\omega_{pll}} \tag{19.36}$$

Observe that if ω_z is allowed to increase indefinitely (i.e. the zero is eliminated from the loop filter making it a pure integration, K_{lp}/s) the poles of (19.35) approach the unit circle; hence, the zero is required to ensure stability of the PLL.

If $\omega_z \gg \omega_{pll}$ the loop will be underdamped resulting in poor tracking behavior. For example, in Fig. 19.9 a large overshoot is observed in the transient step response for the case $Q = 1.8$. This implies greater delay in obtaining lock and larger transient discrepancies between the input and output phase. Furthermore, considerable peaking is observed in the magnitude response; this is called *jitter peaking* and is generally undesirable since noise in the PLL will be amplified within a certain frequency band.

4. For those more familiar with damping factor, ζ, of a second-order transfer function, note that $\zeta = 1/2Q$.



(*a*) (*b*)

Fig. 19.9 Closed-loop PLL responses for the second-order model, $H(s)$, with equivalent bandwidth but different Q: (*a*) time-domain step response; (*b*) frequency-domain jitter transfer function magnitude response

With $Q = 0.5$ although overshoot and jitter peaking are still present, they are greatly reduced.[5] This is a popular design choice, resulting in a closed-loop response with $\omega_z = \omega_{pll}/2$ and coincident real poles at $-\omega_{pll}$.

$$H(s) = \frac{N(1 + s/\omega_z)}{(1 + s/\omega_{pll})^2} \tag{19.37}$$

The 3-dB bandwidth of (19.37) is

$$\omega_{3dB} = 2.5\omega_{pll} \tag{19.38}$$

Key Point: *The design of a second-order PLL with Q = 0.5 provides good transient settling behavior without excessive spreading of the loop filter time constants. A design with Q = 0.1 offers even better tracking behavior, but generally demands a larger loop-filter time constant for the same loop bandwidth.*

Another possibility is to take $Q \ll 0.5$ providing further improvements in transient response and reduced jitter peaking. For example, with $Q = 0.1$ ($\omega_z = 0.1\omega_{pll}$) less than 0.1 dB of jitter peaking is observed in the magnitude response. With $Q \ll 0.5$, the PLL second-order response can be approximated as $1 + s/(\omega_{pll}Q) + s^2/\omega_{pll}^2 \cong (1 + s/\omega_{pll}Q)(1 + sQ/\omega_{pll})$. Since $\omega_0 Q = \omega_{pll}Q = \omega_z$, the first term is equal to $(1 + s/\omega_z)$ which cancels the zero in the numerator of (19.35). With finite $Q \ll 0.5$ the cancellation is approximate and the closed-loop PLL response may be estimated from a first-order model,

$$H(s) \cong \frac{N}{1 + \dfrac{s\omega_z}{\omega_{pll}^2}} \tag{19.39}$$

5. Although an all-pole system with $Q = 0.5$ will have no transient overshoot or peaking in its magnitude response, the closed loop PLL $H(s)$ has a zero which introduces both even when $Q \le 0.5$.

having a 3-dB bandwidth

$$\omega_{3dB} = \omega_{pll}^2/\omega_z = \omega_{pll}/Q \tag{19.40}$$

and a settling time constant

$$\tau \cong \frac{\omega_z}{\omega_{pll}^2} = \frac{Q}{\omega_{pll}} \tag{19.41}$$

Comparing two PLLs with the same closed-loop bandwidth, one with $Q = 0.1$ and the other with $Q = 0.5$, with see that the design with $Q = 0.1$ offers better tracking behavior, but results in a value of ω_z approximately 20 times smaller. This low-frequency zero demands a relatively large time-constant in the loop filter, often requiring off-chip components.

EXAMPLE 19.4

Sketch the bode plot of an open-loop second-order PLL response, $L(s)$. What is the phase margin with $Q = 0.5$ and $Q = 0.1$?

Solution

The open-loop response $L(s)$ in (19.33) has two poles at the origin—one from the loop filter and the other from the VCO whose output phase is the integral of its frequency. It also has one zero at $-\omega_z$. A bode plot is sketched in Fig. 19.10. The phase response is given by $-180° + \text{atan}(\omega/\omega_z)$ and the phase margin is simply $\text{atan}(\omega_t/\omega_z)$. Using (19.33) and solving for $|L(j\omega_t)|^2 = 1$, it may be shown that

$$\omega_t = \frac{\omega_{pll}}{\sqrt{2}} \cdot \sqrt{\frac{1}{Q^2} + \sqrt{\frac{1}{Q^4} + 4}} \tag{19.42}$$

and the phase margin is

$$\text{Phase Margin} = \text{atan}\left(\frac{1}{\sqrt{2}Q} \cdot \sqrt{\frac{1}{Q^2} + \sqrt{\frac{1}{Q^4} + 4}}\right) \tag{19.43}$$

Substitution into (19.43) reveals a phase margin of $76°$ when $Q = 0.5$. With $Q \ll 1$, (19.42) simplifies to $\omega_t \cong \omega_{pll}/Q = \omega_{pll}^2/\omega_z$ and the phase margin becomes

$$\text{Phase Margin} \cong \text{atan}\left(\frac{\omega_{pll}^2}{\omega_z^2}\right) = \text{atan}\left(\frac{1}{Q^2}\right) \tag{19.44}$$

Equation (19.44) reveals a phase margin very close to $90°$ when $Q = 0.1$. Note that in practice, additional poles in the PLL (for example, introduced by the capacitor C_2 in the charge-pump loop filter of Fig. 19.7) will reduce the phase margin from these values.

19.2.2 Limitations of the Second-Order Small-Signal Model

A significant limitation of any continuos-time linear PLL model is that it fails to accurately predict loop behavior at high frequencies. For example, the analysis so far suggests that a PLL can be designed to have an arbitrarily large loop bandwidth. This is untrue. The problem is that a continuous-time model is being applied to a discrete-time

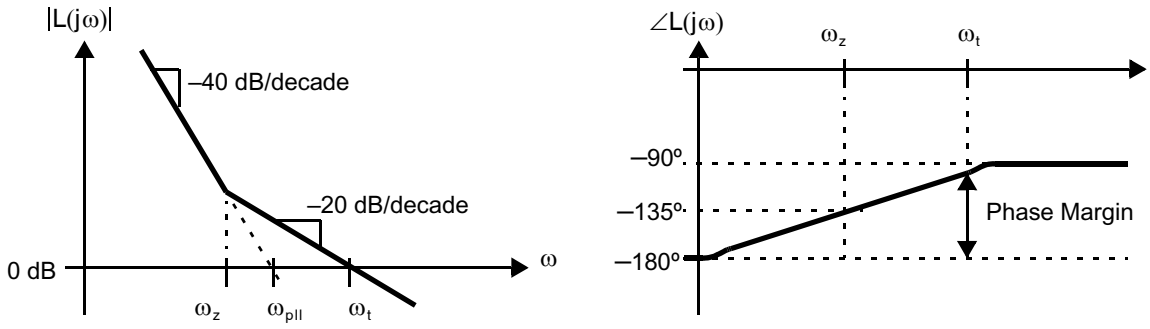


Fig. 19.10 Bode plot of open loop second-order PLL response, $L(s)$.

system. Specifically, the phase detector in Example 19.2 takes periodic snapshots of the phase difference between the divider output and reference input. Hence, it is a discrete-time block operating at the reference frequency. In practice, this means that loop bandwidths are restricted to significantly less than the reference frequency. Intuitively, this is obvious since phase changes can not be tracked faster than they can appear at the input. Several cycles of the reference input must be provided to allow the loop filer to average out the signal from the phase detector and adjust the VCO control input accordingly. Hence, the loop bandwidth is limited to at most roughly $1/20$th to $1/10$th of the reference frequency. This is a significant limitation since we often wish to maximize the bandwidth over which the output clock effectively tracks the reference. Furthermore, higher loop bandwidths permit the use of smaller components in the loop filter.

Key Point: *Due to the discrete-time nature of the PLL, its loop bandwidth is limited by stability considerations to at most roughly 1/20th to 1/10th of the reference frequency, a fact not predicted by the continuous-time linear model.*

EXAMPLE 19.5

A PLL is to operate with a reference clock frequency of 10 MHz and an output frequency of 200 MHz. Design a loop filter for the charge-pump phase comparator so that the loop bandwidth is approximately $1/20$th of the reference clock frequency and $Q = 0.1$. Assume the VCO has $K_{osc} = 2\pi \times 10^7$ rad/Vs and the charge-pump current is $I_{ch} = 20\mu A$.

Solution

From (19.32) and (19.40), it is known that with $Q \ll 0.5$ the loop bandwidth is approximately

$$\omega_{3dB} \cong \frac{\omega_{pll}}{Q} = \frac{1}{Q}\sqrt{\frac{K_{pd}K_{lp}K_{osc}}{N}} \Rightarrow K_{lp} = \frac{Q^2\omega_{3dB}^2 N}{K_{pd}K_{osc}} \tag{19.45}$$

Substituting (19.23) and (19.27) for a charge-pump PLL into (19.45) and rearranging,

$$C_1 = \frac{I_{ch}K_{osc}}{2\pi Q^2\omega_{3dB}^2 N} \tag{19.46}$$

In this example with $\omega_{3dB} = 2\pi \cdot 10$ MHz$/20$, substitution of the known values into (19.46) yields a value very close to $C_1 = 100$ pF. Also from (19.40), with $Q \ll 0.5$ it is known that $\omega_z \cong \omega_{3dB}Q^2$. Hence, using (19.28) for a charge-pump loop filter reveals,

$$R = \frac{1}{C_1 \omega_{3dB} Q^2} \tag{19.47}$$

which in this example yields in a resistor value of $R = 314\ \text{k}\Omega$.

The need for C_2 in the loop filter circuit of Fig. 19.7 is revealed by Example 19.5. Consider what would happen in the absence of C_2 when the charge pump is active; all of the current $I_{ch} = 20\mu\text{A}$ would be driven through the series branch R and C_1 generating a voltage of 6.28 V across the resistor R. Such a large voltage will likely cause some of the transistors in the charge pump to turn off, disturbing the dynamics of the loop.

With C_2 in place as shown in Fig. 19.7, these large voltage transients will be suppressed since the voltage across C_2 can not change instantaneously. The resulting loop filter transfer function is repeated here from (19.25).

$$K_{lp}H_{lp}(s) = \frac{1}{s(C_1 + C_2)} \cdot \frac{1 + sRC_1}{1 + sR\left(\frac{C_1 C_2}{C_1 + C_2}\right)} \tag{19.48}$$

Comparing (19.48) and (19.26) reveals that the addition of C_2 has two effects:

- K_{lp} is decreased, but this effect will be small if C_2 is taken much smaller than C_1
- a real pole is introduced at a frequency $(C_1 + C_2)/(RC_1C_2) > \omega_z$

The additional pole makes the closed-loop system third-order and reduces the phase margin somewhat compared to the results obtained in Example 19.4 for a second-order loop. However, placing the new pole at approximately 2 times the loop bandwidth, $2\omega_{3dB}$, provides the desired glitch-suppression without much effect on the settling time or jitter peaking of the loop. For example, with $Q = 0.5$ this corresponds to selecting C_2 equal to one-tenth to one-eighth the value of C_1. With $Q = 0.1$, the value of C_2 should be roughly $1/250$th the value of C_1.

EXAMPLE 19.6

Select a capacitor value C_2 for the loop filter designed in Example 19.5.

Solution

Since the loop is designed with $Q = 0.1$, a value of $C_2 = C_1/250 \cong 0.4$ pF is chosen.

Example 19.5 and Example 19.6 also illustrate the practical challenge of designing for a very small value of Q: it leads to large component values and large spreads in component values.

EXAMPLE 19.7

Repeat the design in Example 19.5 but with $Q = 0.5$.

Solution

The design is performed by initially using the second-order model and ignoring C_2. Once C_1 and R are chosen, a value for C_2 is calculated under the assumption that it will have only a small effect on the loop bandwidth. For $Q = 0.5$, using (19.32) and (19.38),

$$\omega_{3dB} \cong 2.5\omega_{pll} = 2.5\sqrt{\frac{K_{pd}K_{lp}K_{osc}}{N}}$$
$$\Rightarrow K_{lp} = \frac{\omega_{3dB}^2 N}{(2.5)^2 K_{pd}K_{osc}} \tag{19.49}$$

Again substituting (19.23) and (19.27) for a charge-pump PLL into (19.49) and rearranging,

$$C_1 = \frac{(2.5)^2 I_{ch}K_{osc}}{2\pi\omega_{3dB}^2 N} \tag{19.50}$$

The known values, including $\omega_{3dB} = 2\pi \cdot 10\ \text{MHz}/20$, substituted into (19.50) yields a capacitance of $C_1 = 6.25$ pF. Combining (19.38) and $Q = \omega_z/\omega_{pll}$, we see that $\omega_z = Q\omega_{3dB}/2.5$ which, for a charge-pump loop filter, implies

$$R = \frac{2.5}{C_1\omega_{3dB}Q} \tag{19.51}$$

The numerical result in this case is $R = 255\ \text{k}\Omega$. Finally, a value of $C_2 = C_1/10 = 0.625$ fF is used to suppress glitches while placing the additional pole beyond the loop bandwidth.

Notice that the component values in Example 19.7 with $Q = 0.5$ are much more readily integrated than those in Example 19.5 and Example 19.6 which use $Q = 0.1$. The trade-off is increased overshoot and jitter peaking in the closed-loop PLL response.

19.2.3 PLL Design Example

Although the details of a PLL design procedure will depend, of course, on which aspects of the system are specified, the following points provide basic direction.

- The ratio between the desired output frequency and the input reference frequency sets the divider ratio, N.
- The VCO gain constant K_{osc} is usually determined by circuit design considerations in the VCO, but generally must be large enough to allow the VCO to cover the expected variations in its free-running frequency and desired output frequency.
- When minimal jitter peaking or fast settling of the loop are important design specifications, a value of $Q \ll 0.5$ is desired, with $Q = 0.1$ being a reasonable choice. However, $Q = 0.5$ will usually result in loop filer component values that are easier to realize.
- When a high-quality (low-noise) input reference clock is available, a high loop bandwidth is desirable. However, the loop bandwidth is limited to at most one-tenth of the input reference frequency to ensure robust stability.
- Once the loop bandwidth is specified, choose $\omega_{pll} = 0.4\omega_{3dB}$ for $Q = 0.5$ [6], or take $\omega_{pll} = Q\omega_{3dB}$ for $Q \ll 0.5$ [7].
- With N and K_{osc} determined as described, K_{pd} and K_{lp} are chosen to give the desired value of $\omega_{pll} = \sqrt{K_{pd}K_{lp}K_{osc}/N}$ [8]. In a charge-pump PLL, this amounts to choosing $I_{ch} = 2\pi K_{pd}$ [9] and C_1 ($\cong 1/K_{lp}$ neglecting C_2 [10]). The designer has one degree of freedom here which is used to influence the

---

6. equation (19.38)
7. equation (19.40)
8. equation (19.32)
9. equation (19.23)
10. equation (19.27)

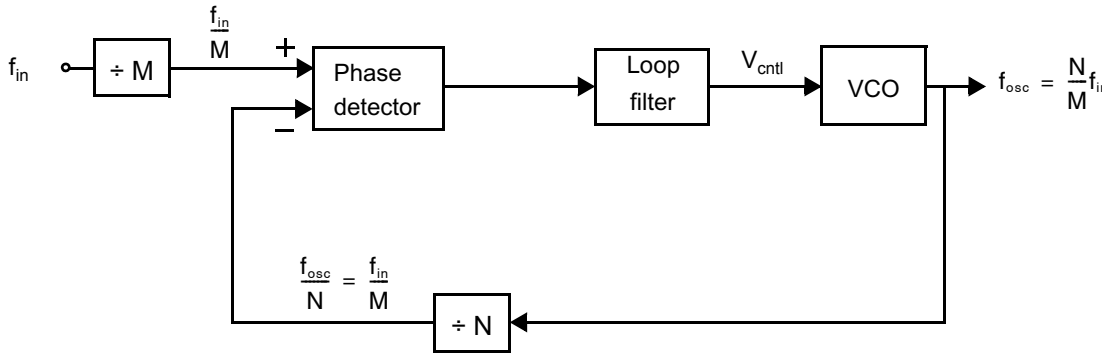


Fig. 19.11 A general integer-N PLL architecture.

size of the loop filter components, the power consumption in the phase detector, and the noise in the loop filter (since a smaller resistor value will generally reduce noise in the loop filter).

- Take $\omega_z = Q\omega_{pll}$ [11]. In a charge-pump PLL, this amounts to choosing $R = 1/\omega_z C_1$ [12].
- In a charge-pump PLL, an additional pole is introduced by the glitch-suppression capacitor C_2. With $Q = 0.5$, select C_2 between one-tenth and one-eighth the value of C_1. With $Q = 0.1$, the value of C_2 should be roughly $1/250$th the value of C_1.

EXAMPLE 19.8

Design a charge-pump PLL that accepts a fixed 40 MHz reference and can produce an output clock frequency anywhere in the range 1.4 GHz to 1.6 GHz in 20 MHz steps. For each output frequency, a different division ratio is used. In order to cover the full range of output frequencies while the control voltage changes by only 2 V, K_{osc} is chosen to be $2\pi \cdot 10^8$ rad/s · V. Select the value of I_{ch} and the loop filter components C_1, R, and C_2.

Solution

The output frequency is N-times higher than the input frequency, so changing the division ratio N allows the output frequency to be placed at different integer multiples of the input reference. Hence, this architecture is called an ***integer-N PLL***. When the output frequency must be set with a resolution finer than the available reference signal, an additional divider is applied to the input to reduce its frequency. The general architecture is shown in Fig. 19.11 and produces an output frequency of $(N/M)f_{in}$.

In this case, since the output frequency resolution requirement is 20 MHz, a divide-by-2 ($M = 2$) is applied to the 40-MHz input. To cover the full range of output frequencies, N must be programmable to any integer from 70 to 80. A value of $Q = 0.5$ is selected and, assuming the input reference has very low noise, the loop bandwidth is set to one-tenth of 20 MHz: $\omega_{3dB} = 2\pi \cdot 2$ MHz. Hence,

$$\omega_{pll} = 0.4\omega_{3dB} = 2\pi \cdot 800 \text{ kHz}$$

11. equation (19.36)
12. equation (19.28)

Combining equations (19.23), (19.27), and (19.32) and rearranging yields

$$\frac{I_{ch}}{C_1} = \omega_{pll}^2 \frac{2\pi N}{K_{osc}} \tag{19.52}$$

Substituting the values in this example, $I_{ch}/C_1 \cong 20 \cdot 10^6$ V/s. A reasonable choice is $I_{ch} = 100\ \mu A$ and $C_1 = 5$ pF. The resistor must then be chosen to place a zero at $\omega_z = Q\omega_{pll} = 2\pi \cdot 400$ kHz. Hence, $R = 1/(\omega_z C_1) \cong 80\ k\Omega$. Finally, with $Q = 0.5$, a good choice is to select $C_2 = C_1/10 = 0.5$ pF.

19.3 JITTER AND PHASE NOISE

Key Point: *Jitter is a measure of random variations in the timing phase of a clock signal and phase noise is its frequency-domain representation. They present fundamental limitations on the accuracy of time references.*

Key Point: *The absolute jitter of a clock, τ_k, is a discrete-time sequence giving the displacement of its transition times away from their ideal values.*

Unlike most analog signals on integrated circuits where the value of a voltage or current bears the information, the important aspects of a clock waveform are the times at which it transitions across some particular threshold. Jitter is the random variation in these times and phase noise is the frequency-domain representation of jitter. Since jitter and phase noise arise from thermal and other noise processes present in all circuits, they represent fundamental limitations on the accuracy of any on-chip time reference making them central to PLL design.

A typical clock signal is illustrated in Fig. 19.12. The transition times are labeled t_k. These may be the times when a logic signal changes state or when current switches from one branch to another. For this example only the rising transitions are considered, but in a real circuit either the rising transitions, falling transitions, or both may be of interest. For an ideal clock the transitions are equally spaced so that $t_k = kT_0$ where T_0 is the *clock period*. In any practical clock, however, the transitions are displaced from their ideal times.

Specifically, let the deviation in the transition times from their ideal values be a discrete-time sequence τ_k.

$$t_k = kT_0 + \tau_k. \tag{19.53}$$

The sequence τ_k is the *absolute jitter* of the clock in time units. It can also be normalized to radians.

$$\phi_k = \tau_k \cdot (2\pi/T_0) \tag{19.54}$$

Absolute jitter is a useful analytical tool. However, it is not directly measured in a lab since doing so requires access to an "ideal" clock.

EXAMPLE 19.9

Find the absolute jitter of a sinusoid of amplitude A and angular frequency ω_0 in additive white noise $n(t)$ with variance σ_n^2.

Solution

The sinusoid may be described as follows:

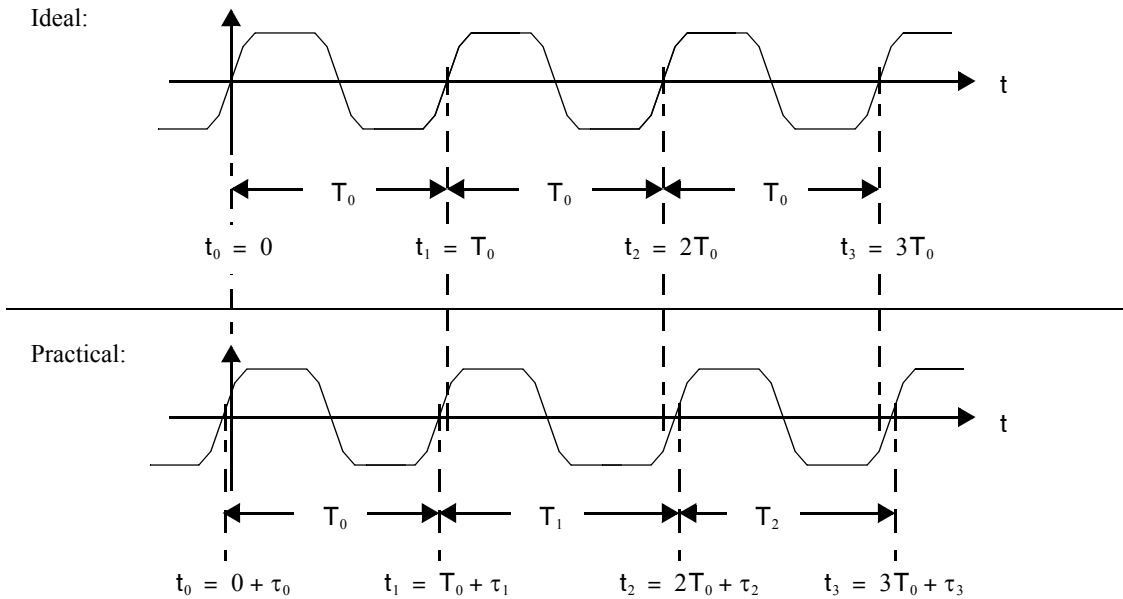


Fig. 19.12 Illustration of absolute jitter in a practical clock waveform.

$$v(t) = A\sin(\omega_0 t) + n(t) \tag{19.55}$$

Assuming the noise n is much smaller than the amplitude of oscillation, $\sigma_n \ll A$, the transition times of $v(t)$ are disturbed only slightly from their nominal values kT_0. The amount of disturbance can be estimated from the slope of the sinusoid around the transition times, $A\omega_0$, and samples of the noise, $n(t)$, labeled n_k.

$$\tau_k \cong \frac{n_k}{A\omega_0} \tag{19.56}$$

The discrete-time noise process, n_k, is white with variance σ_n^2. Hence, absolute jitter is also a discrete-time white noise process with variance

$$\sigma_\tau^2 = \frac{\sigma_n^2}{A^2\omega_0^2} \tag{19.57}$$

At high frequencies ω_0 and when the sinusoid has high amplitude A, the jitter is reduced because the sinusoid will have a high slope around its zero-crossings. Expressed in radians, the effect of frequency is cancelled.

$$\sigma_\phi^2 = \sigma_\tau^2 \cdot \left(\frac{2\pi}{T_0}\right)^2 = \frac{\sigma_n^2}{A^2} \tag{19.58}$$

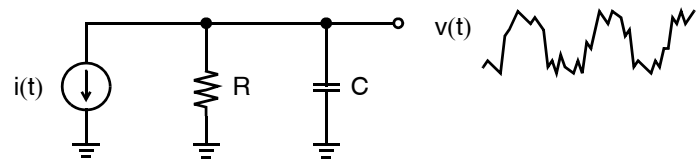


Fig. 19.13 A noise source resistance introduces jitter.

EXAMPLE 19.10

In Fig. 19.13, assume $i(t)$ is an ideal sinusoidal current source with amplitude 1 mA, $R = 1k\Omega$, and $C = 1pF$. Find the variance of absolute jitter at $v(t)$ when the signal is at a frequency 1 MHz. Also find the variance when the signal is at 1 GHz.

Solution

The voltage $v(t)$ will be in exactly the form of (19.55) where $n(t)$ is the gaussian-distributed voltage noise arising from the thermal noise of R. The variance of $n(t)$ is

$$\sigma_n^2 = \frac{kT}{C} \tag{19.59}$$

which, at a temperature of $300K$ has an rms value of $64.4\mu V$. The amplitude, A, of the sinusoid in (19.55) depends upon the oscillation frequency f. At 1 MHz, the amplitude will be very close to $(1mA)(1k\Omega) = 1V$. Hence, by substitution into (19.57),

$$\sigma_\tau^2 = \frac{(64.4\mu V)^2}{(1V)^2(2\pi 10^6)^2} = (10.2ps)^2$$

At 1 GHz, the amplitude at $v(t)$ is attenuated by the lowpass RC filter so that $A = 0.157V$. Hence,

$$\sigma_\tau^2 = \frac{(64.4\mu V)^2}{(0.157V)^2(2\pi 10^9)^2} = (0.065ps)^2$$

Note that although in time units there is less jitter in the 1 GHz case, the jitter expressed in radians is actually higher at 1 GHz because of the reduced amplitude at $v(t)$.

Phase noise is the frequency-domain representation of absolute jitter. Thus, we define $S_\phi(f)$, the phase noise of a clock signal with random jitter, as the power spectral density of the sequence ϕ_k.[13] Recall that a voltage noise spectral density specifies the power of a voltage signal in any 1-Hz-wide bin of frequency spectrum, and hence has units of V^2/Hz. Similarly, $S_\phi(f)$ specifies the power of ϕ_k in rad^2 within any 1-Hz-wide frequency bin, and hence has units of rad^2/Hz. Furthermore, the variance of ϕ_k is given by the integral of $S_\phi(f)$ over all frequencies. Since it is customary to quantify jitter in units of time rather than radians, this is usually expressed as follows:

13. Specifically, $S_\phi(f)$ is the one-sided power spectral density of ϕ_k, defined over the range $0 \le f \le 1/2T_0$ and obtained from the discrete-time Fourier transform of the autocorrelation of ϕ_k.

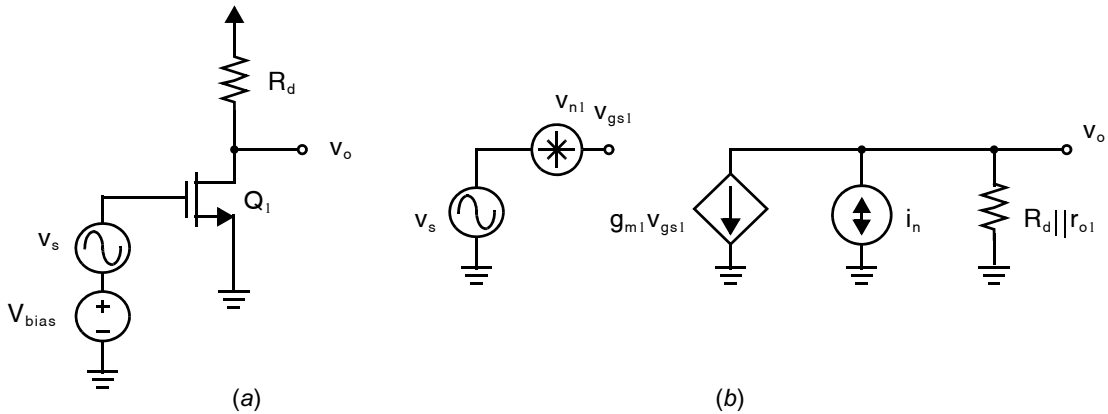


Fig. 19.14 (*a*) Common source amplifier in Example 19.11. (*b*) Small-signal equivalent, including transistor noise sources.

$$\sigma_\tau^2 = \left(\frac{T_0}{2\pi}\right)^2 \int_0^{1/2T_0} S_\phi(f)df \tag{19.60}$$

Taking the square root of both sides gives the root-mean-square (rms) absolute jitter.

$$\sigma_\tau = \frac{T_0}{2\pi}\sqrt{\int_0^{1/2T_0} S_\phi(f)df} \tag{19.61}$$

Equations (19.60) and (19.61) are key relationships between absolute jitter in the time-domain and phase noise.

EXAMPLE 19.11

Consider the common source amplifier in Fig. 19.14(*a*). Transistor Q_1 is biased in saturation by dc source V_{bias} and v_s is a noiseless small-signal sinusoid of amplitude $|V_s|$. Estimate the phase noise at the output including thermal noise of the resistor and transistor as well as flicker noise.

Solution

The small-signal equivalent circuit is shown in Fig. 19.14(*b*). The voltage noise source v_{n1} models the flicker noise of the transistor Q_1 and has a noise spectral density,

$$V_{n1}^2(f) = \frac{K}{WLC_{ox}f}$$

The current noise source i_n includes the thermal noise of both the transistor and resistor,

$$i_n^2(f) = 4kT\left(\frac{2}{3}\right)g_m + \frac{4kT}{R_d}$$

The resulting voltage at v_0 takes on the form of (19.55)—a sinusoidal voltage in additive noise. The sinusoidal component has amplitude $|V_s|g_{m1}(R_d||r_{o1})$. The voltage noise spectral density at v_0 is obtained by a routine noise circuit analysis of Fig. 19.14(*b*).

$$v_{o,n}^2(f) = \left[\frac{K}{WLC_{ox}f}g_{m1}^2(R_d||r_{o1})^2 + 4kT\left(\frac{2}{3}\right)g_m(R_d||r_{o1})^2 + \frac{4kT}{R_d}(R_d||r_{o1})^2\right]$$

Assuming that the zero-crossings of the sinusoidal waveform at v_0 are disturbed only a small amount by the noise, the approximation in (19.56) may be applied. In terms of phase,

$$\phi_k \cong \frac{n_k}{|V_s|}$$

where n_k is the sampled noise voltage $v_{o,n}$. Therefore, the phase noise is,

$$S_\phi(f) = \frac{(R_d||r_{o1})^2}{|V_s|^2}\left[\frac{K}{WLC_{ox}f}g_{m1}^2 + 4kT\left(\frac{2}{3}\right)g_m + \frac{4kT}{R_d}\right]$$

Note that it has a $1/f$ component and a white component.[14]

19.3.1 Period Jitter

Several other measures of jitter are also possible. A jittered clock will have a time-varying period, as shown in Fig. 19.12,

$$T_k = t_{k+1} - t_k = T_0 + \tau_{k+1} - \tau_k \tag{19.62}$$

The deviation of the period away from its nominal value of T_0 is the ***period jitter***, also sometimes referred to as the cycle jitter.

$$J_k = T_k - T_0 = \tau_{k+1} - \tau_k \tag{19.63}$$

Equation (19.63) reveals that period jitter is obtained by passing the absolute jitter sequence, τ_k, through a discrete-time $(z-1)$ filter. Hence, its variance is the integral of phase noise filtered by a $(z-1)$ filter.

$$\sigma_J^2 = \left(\frac{T_0}{\pi}\right)^2 \int_0^{1/2T_0} \sin^2(\pi f T_0) S_\phi(f) df \tag{19.64}$$

The term $\sin^2(\pi f T_0)$ in the integrand of (19.64) is due to the $(z-1)$ filter.

Period jitter is more easily measured in a lab than the absolute jitter, since no reference is required. Furthermore, the period jitter is of significant practical importance in digital systems where the maximum permissible logic delays will depend on the period jitter of the synchronizing clock.

14. Strictly speaking, ϕ_k is a discrete-time random process; hence, there may be some aliasing of the white noise in $v_{o,n}$ causing the phase noise to be larger than this. Naturally, the white noise is assumed to be eventually bandlimited by some capacitance at v_0 ensuring it has finite power.

19.3.2 P-Cycle Jitter

The idea of period jitter is easily extended by considering variations in the duration of P consecutive clock cycles from its ideal value, PT_0.

$$J(P)_k = t_{k+P} - t_k - PT_0 = \tau_{k+P} - \tau_k \tag{19.65}$$

Called the P-cycle period jitter or simply *P-cycle jitter*, $J(P)_k$ is obtained by passing the absolute jitter sequence, τ_k, through a discrete-time $(z^P - 1)$ filter. Hence, its variance is the integral of phase noise filtered by a $(z^P - 1)$ filter.

$$\sigma^2_{J(P)} = \left(\frac{T_0}{\pi}\right)^2 \int_0^{1/2T_0} \sin^2(\pi f P T_0) S_\phi(f) df \tag{19.66}$$

Period jitter is a special case of the P-cycle jitter where $P = 1$. Measurements of $\sigma_{J(P)}$ for different values of P are easily done in a lab and reveal much information about the spectral properties of phase noise due to the differently shaped filters applied in (19.66) for different values of P.

19.3.3 Adjacent Period Jitter

Yet another measurement of jitter is obtained by looking at the difference between the duration of each period and the following period.

$$J_{k+1} - J_k \tag{19.67}$$

The so-called *adjacent period jitter*, also sometimes called cycle-to-cycle jitter, is therefore mathematically obtained by passing the period jitter J_k through yet another $(z - 1)$ filter or equivalently by passing the absolute jitter through a $(z - 1)^2$ filter. Hence, it is related to phase noise by the following integral,

$$\sigma^2_c = \left(\frac{2T_0}{\pi}\right)^2 \int_0^{1/2T_0} \sin^4(\pi f T_0) S_\phi(f) df \tag{19.68}$$

EXAMPLE 19.12

What is the rms period jitter of the sinusoid in additive white noise from Example 19.9?

Solution

Again assuming the noise is much smaller than the amplitude of oscillation, $\sigma_n \ll A$, it was shown in Example 19.9 that the absolute jitter is also white. Hence, the phase noise is constant. Combining (19.57) and (19.60) therefore yields

$$S_\phi(f) = \left(\frac{8\pi^2}{T_0}\right)\left(\frac{\sigma_n^2}{A^2\omega_0^2}\right) = \frac{2T_0\sigma_n^2}{A^2} \tag{19.69}$$

Integration in (19.64) gives the variance of the period jitter.

$$\sigma_J^2 = \frac{2\sigma_n^2}{A^2\omega_0{}^2} \tag{19.70}$$

The rms period jitter is the square root of (19.70).

EXAMPLE 19.13

A clock's zero crossings are modulated by a sinusoidal disturbance at a frequency f_m so that

$$\phi_k = \Phi_0 \sin(2\pi f_m t) \tag{19.71}$$

What is the resulting rms absolute jitter, period jitter, and adjacent period jitter?

Solution

The phase noise $S_\phi^2(f)$ will have a discrete spectral tone at f_m.

$$S_\phi(f) = \frac{\Phi_0^2}{2}\delta(f - f_m)$$

Key Point: *Periodic variations in the phase of a clock result in discrete tones in phase noise spectra and are called spurs.*

Discrete tones in phase noise spectra are called *spurs* because on a plot of $S_\phi^2(f)$ versus frequency they appear as vertical spikes. The integral in (19.60) is easily evaluated to give the absolute jitter,

$$\sigma_\tau^2 = \left(\frac{T_0}{2\pi}\right)^2\left(\frac{\Phi_0^2}{2}\right) = \frac{1}{8}\left(\frac{T_0\Phi_0}{\pi}\right)^2 \tag{19.72}$$

The period jitter is reduced slightly by the extra term in the integrand of (19.64) because the change in absolute jitter from one transition to the next is less than the total range over which it deviates.

$$\sigma_J^2 = \frac{1}{2}\left(\frac{T_0\Phi_0}{\pi}\right)^2 \sin^2(\pi f_m T_0) \tag{19.73}$$

Note that if the phase modulation is slow (i.e. $f_m T_0 \ll 1$) the period jitter is small since ϕ_k varies very little from one cycle to the next. The same is true of adjacent period jitter.

$$\sigma_C^2 = 2\left(\frac{T_0\Phi_0}{\pi}\right)^2 \sin^4(\pi f_m T_0) \tag{19.74}$$

19.3.4 Other Spectral Representations of Jitter

Phase noise is the spectral representation of the phase of an oscillatory waveform, but this is difficult to observe in practice since it requires a perfect time-reference. It is much easier to observe the spectrum of the oscillatory waveform itself. Let us denote the power spectral density of a clock waveform (e.g. $v(t)$ in Example 19.10) $S_v(f)$.

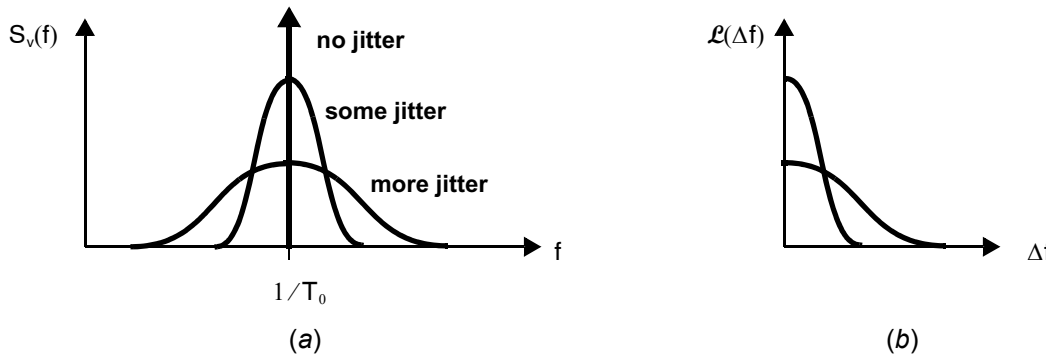


Fig. 19.15 (*a*) The power spectral densities of oscillatory waveforms with and without jitter. (*b*) The normalized power spectral density, $\mathcal{L}(\Delta f)$.

Clearly it will be concentrated around the oscillating frequency, ω_0, sometimes called the *carrier frequency*. A pure sinusoid with no variation it its period (hence, no jitter) will of course have only a single tone at $(1/T_0)$. A practical signal, however, will have power in a range of frequencies around the carrier, as shown in Fig. 19.15.

Signals with more jitter will have more power in the tails of $S_v(f)$. However, the power in those tails is also proportional to the squared amplitude of the oscillating waveform. Hence, to obtain a measure of jitter, $S_v(f)$ may be normalized with respect to the total power in the signal, which for a sinusoid is $A^2/2$. Shifting the resulting spectrum down to dc gives the following alternative spectral representation of jitter:

$$\mathcal{L}(\Delta f) = \frac{S_v\left(\frac{1}{T_0} + \Delta f\right)}{A^2/2} \tag{19.75}$$

The units of $\mathcal{L}(\Delta f)$ are Hz^{-1}, but it is most common to plot its logarithm $10\log_{10}(\mathcal{L}(\Delta f))$ in units of dBc/Hz where the "c" denotes that $\mathcal{L}(\Delta f)$ has been normalized with respect to the total power in the "carrier." The argument Δf is called the *offset frequency* since $\mathcal{L}(\Delta f)$ represents the signal spectrum at a frequency offset from the carrier by Δf Hz.

Strictly speaking, the relationship between $\mathcal{L}(\Delta f)$ and $S_\phi(f)$ is very complex but the following simple approximation may be applied to spectral components where $S_\phi(f)$ is very small.[15]

$$\mathcal{L}(\Delta f) \approx \frac{S_\phi(\Delta f)}{2}, \qquad S_\phi(\Delta f) \ll 1 \tag{19.76}$$

The factor of two in (19.76) arises because the definition of $\mathcal{L}(\Delta f)$ in (19.75) ignores one of the sidebands of $S_v(f)$. We shall see that practical oscillators always have very large phase noise at small offset frequencies, as suggested by Fig. 19.15. Hence, (19.76) is only applied to large offset frequencies where it becomes a reasonable approximation. Combining (19.75) and (19.76) allows one to estimate the phase noise of a signal by looking only at its spectrum. Because of this equality, phase noise plots of $S_\phi(f)$ are often also shown logarithmically with units labelled dBc/Hz, although its units are more accurately written $10\log_{10}(rad^2/Hz)$.

15. Each component of $S_\phi(f)$ gives rise to spectral content at a multiplicity of offset frequencies in $\mathcal{L}(\Delta f)$, the amplitudes of which are related by Bessel functions. If the phase modulation of the carrier is very small, a linear approximation is possible that ignores all but the first (largest) sideband in $\mathcal{L}(\Delta f)$ arising from each component of $S_\phi(f)$.

19.3.5 Probability Density Function of Jitter

When designing a digital system the critical path must have a delay shorter than the smallest possible clock period. Hence, the probability density function (PDF) of τ_k and its derivatives are also important measures of clock purity. Note that spectral representations of jitter can not generally be used to estimate the PDF of τ_k.

EXAMPLE 19.14

Find the PDF of τ_k for a sinusoid with amplitude A and angular frequency ω_0 in additive white gaussian noise, $n(t)$, with variance σ_n^2.

Solution

The sinusoid may be described as:

$$v(t) = A\sin(\omega_0 t) + n(t) \tag{19.77}$$

As in Example 19.9, assuming the noise n is much smaller than the amplitude of oscillation, $\sigma_n \ll A$, the absolute jitter is linearly related to n_k by the slope of the sinusoid around its zero crossings.

$$\tau_k \cong \frac{n_k}{A\omega_0}$$

Since n_k is gaussian-distributed, absolute jitter is also gaussian-distributed with the following PDF:

$$f_\tau(t) = \frac{1}{\sigma_\tau\sqrt{2\pi}}\exp\left(-\frac{t^2}{2\sigma_\tau^2}\right) \tag{19.78}$$

where σ_τ is given by (19.57).

If the absolute jitter τ_k has a Gaussian distribution, the period jitter, P-cycle jitter, and adjacent period jitter are all also Gaussian-distributed since they are linear derivatives of τ_k given by (19.63), (19.65) and (19.67).

EXAMPLE 19.15

It is desired to drive a digital system with a clock period that is always at least 1 ns. The clock has a nominal period $T_0 = 1.1\,\text{ns}$ and a gaussian jitter distribution. If the rms period jitter is 5 ps, with what probability will any particular clock period be less than 1 ns?

Solution

The clock period will be less than 1 ns whenever the period jitter $J_k < -0.1$ ns. The PDF of J_k is,

$$f_J(t) = \frac{1}{(5\cdot 10^{-12})\sqrt{2\pi}}\exp\left(-\frac{t^2}{5\cdot 10^{-25}}\right)$$

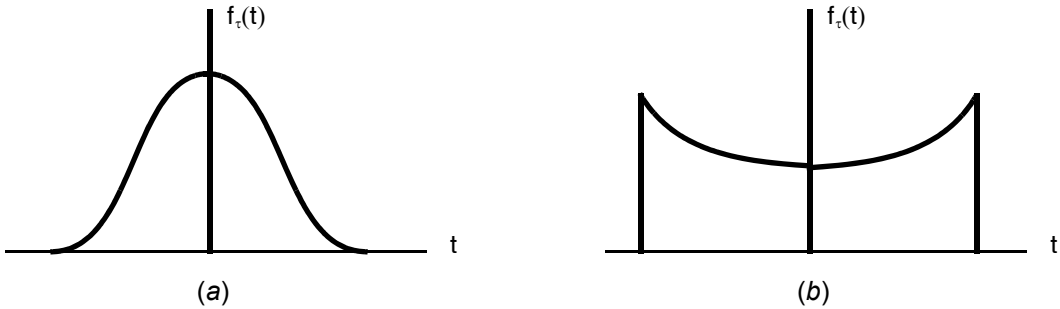


Fig. 19.16 Jitter PDFs: (*a*) Gaussian distribution; (*b*) sinusoidal jitter distribution.

Hence, the probability that the clock period is less than 1 ns is given by the integral of $f_J(t)$ over the range $-\infty$ to -100ps.

$$\int_{-\infty}^{-10^{-10}\,s} f_J(t) \cdot dt = 2.75 \cdot 10^{-89}$$

Gaussian distributions are common when jitter arises due to random noise. This is because noise processes are generally the superposition of many independent noise events, hence the law of large numbers tends to give the resulting jitter an approximately gaussian distribution. However, other jitter distributions arise when outside interference is a major contributor. For example, Fig. 19.16 compares a gaussian jitter distribution to the jitter PDF observed in the presence of sinusoidal phase modulation. In practice, local variations in supply voltage, digital data patterns, and electromagnetic interference from neighboring signals are often superimposed on random circuit noise to produce a complex PDF. Measurement equipment may then be able to decompose the observed jitter PDF into its constituent parts providing clues to the underlying sources of jitter.

19.4 ELECTRONIC OSCILLATORS

In this section, the basics of electronic oscillators are presented. Oscillators come in a variety of different types, but the two major classifications are those that directly create sinusoidal outputs as opposed to those with square (or triangular) wave outputs. *Sinusoidal-output oscillators* are usually realized using some kind of frequency-selective or tuned circuit in a feedback configuration, whereas square-wave-output oscillators are usually realized using a nonlinear feedback circuit such as a relaxation oscillator or a ring counter. Sinusoidal oscillators realized using tuned circuits can be further classified into RC circuits, switched-capacitor circuits, LC circuits, and crystal oscillators. A table showing the classification of oscillators is shown in Fig. 19.17.

The two most common oscillators implemented on integrated circuits are ring oscillators and LC oscillators. Unless precise off-chip components are included in the oscillator circuits, their frequency of operation is poorly-controlled in the presence of process, temperature, and supply voltage variations. Hence, integrated circuit oscillators are nearly always designed with a mechanism for tuning their operating frequency. The actuator used for tuning is generally the voltage at some node in the oscillator circuit, so these circuits are called voltage-controlled oscillators (VCOs).

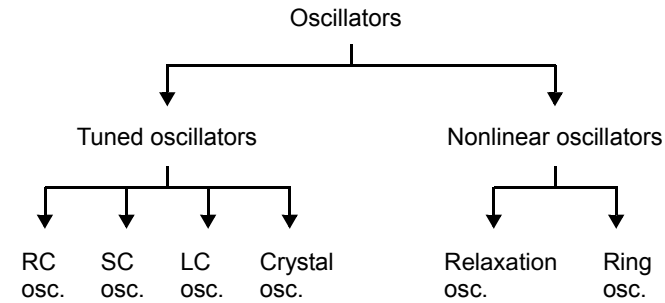


Fig. 19.17 Classification of oscillators.

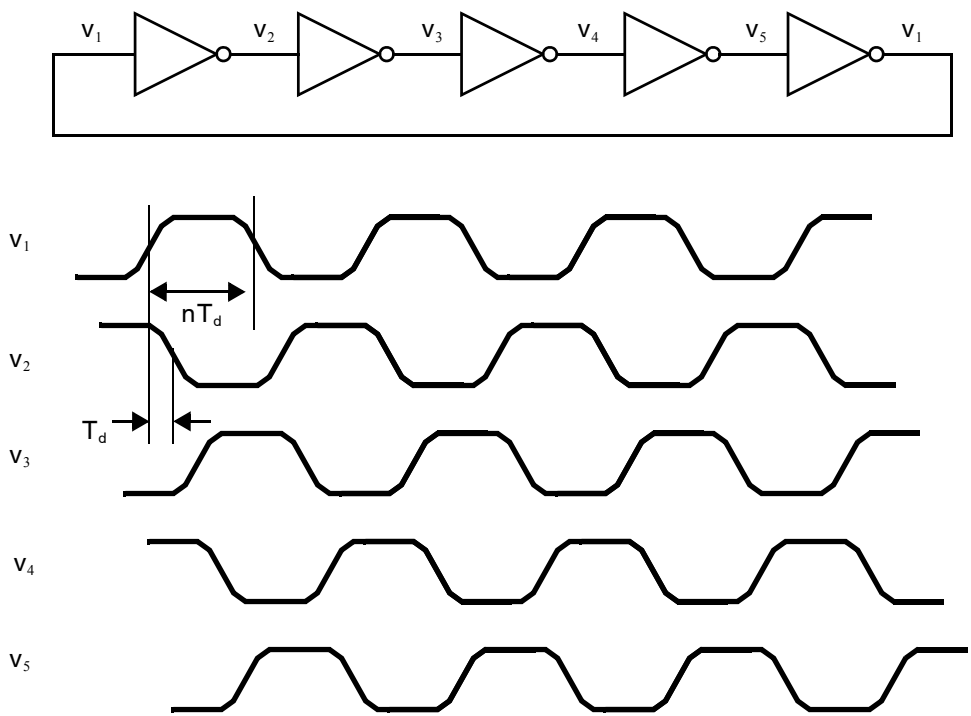


Fig. 19.18 Ring oscillator with n = 5.

19.4.1 Ring Oscillators

One of the most popular ways of realizing digital-output MOS VCOs is to use a *ring oscillator* and add voltage control to it. A ring oscillator is realized by placing an odd number of open-loop inverting amplifiers in a feedback loop. The simplest type of amplifier that can be used is a simple digital inverter, as shown in Fig. 19.18.

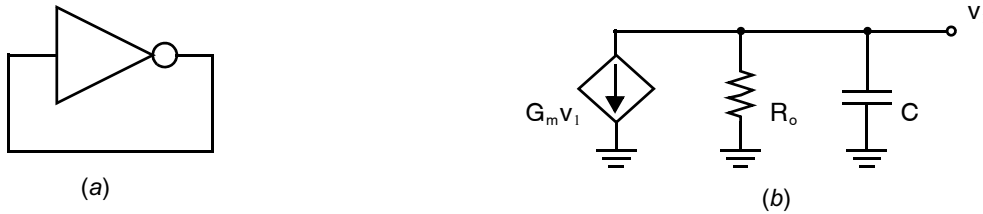


Fig. 19.19 (*a*) A single inverting amplifier, $n = 1$, with its input tied to its output. (*b*) A corresponding small-signal model.

Imagine that, at start-up, some voltage in the circuit transitions from a low-voltage to a high-voltage state. Each half-period, the transition will propagate around the loop with an inversion. For example, assume the output of the first inverter changes to 1. This change will propagate through all five inverters in a time of $T_0/2$, at which time the output of the first inverter will change to 0; after an additional time of $T_0/2$, the first inverter's output will change back to 1, and so on. Assuming each inverter has a delay of T_d and that there are n inverters, the half-period of oscillation is

$$\frac{T_0}{2} = nT_d \tag{19.79}$$

Thus,

$$f_0 = \frac{1}{T_0} = \frac{1}{2nT_d} \tag{19.80}$$

Fig. 19.18 illustrates the case $n = 5$. By making the delay of the inverters voltage controlled, a VCO can be realized.

EXAMPLE 19.16

What is the behavior of a ring with only $n = 1$ inverter, as shown in Fig. 19.19(*a*)?

Solution

Based upon (19.80) and the preceding discussion, one might think that the circuit will oscillate at a frequency $f_0 = 1/(2T_d)$. But the first-order small-signal model of an inverting amplifier in Fig. 19.19(*b*) is clearly an undriven single-time constant circuit whose solution is an exponential decay,

$$v_1(t) = v_1(0)\exp\left(-\frac{t}{R_{o1}C_1}\right)$$

When the inverting amplifier is a simple digital CMOS inverter, this analysis is accurate and there are no oscillations, as one may readily verify in the lab. The voltage quickly settles to a dc value between ground and the supply. In fact, this can be a useful bias circuit for establishing a dc voltage precisely equal to the trip point of the inverter, where $v_{out} = v_{in}$.

If the inverting amplifier has other internal nodes not represented by the small-signal schematic in Fig. 19.19(*b*), it will have a higher-order response and the solution is more complex.

Fig. 19.20 The two stable states of a CMOS latch, $n = 2$.

Example 19.16 illustrates a shortcoming of the simple time-domain interpretation of ring oscillators illustrated in Fig. 19.18. A ring oscillator is actually a feedback circuit, and in order for it to exhibit oscillations it should be unstable.

Applying the first-order model of the inverting amplifier in Fig. 19.19(*b*) to a general ring oscillator with n identical stages, the loop gain can be found by breaking the loop at any point:

$$L(s) = \left(\frac{G_m R_o}{1 + sR_oC}\right)^n \tag{19.81}$$

Assuming $G_mR_o \gg 1$, each inverter has approximately 90° phase shift at its unity-gain frequency. Hence, as long as there are 3 or more inverters, it is guaranteed that the loop gain will still be greater than unity when the phase shift around the loop becomes greater than 180°. But, with $n = 1$ there is only 90° phase shift at the loop's unity-gain frequency so the system is stable.

EXAMPLE 19.17

What is the behavior of a ring of inverters with $n = 2$?

Solution

It is a bistable circuit meaning it has two stable operating points as shown in Fig. 19.20. This is desirable behavior when used as a basic digital latch. No other operating points are stable, so immediately upon startup the circuit will settle into one of these two states and rest there until forced into some other state. The circuit will not oscillate.

Positive-feedback loops as in Example 19.17 are also unstable but do not oscillate. They can be thought of as negative feedback loops with an additional 180° phase shift at dc. Hence, the node voltages simply diverge exponentially until the stages are saturated resulting in a latch-up of the ring. Therefore, if single-ended inverting amplifiers are used, it is required that n is odd.

In many integrated ring oscillators, fully differential inverters are used to obtain better power-supply insensitivity. When this is done, an even number of inverters can be used and the inversion required around the loop can be achieved by simply interchanging the outputs of the last inverter before feeding them back to the input, as shown in Fig. 19.21. This has a very important advantage: the outputs of the middle inverters will have a quadrature phase relationship compared with the last inverters (assuming all inverters and their loads are carefully matched) [Buchwald, 1991]. These quadrature outputs are very useful in many communication applications such as quadrature modulators and some clock-extraction circuits.

An example of realizing a fully differential inverter with a programmable delay is shown in Fig. 19.22, where the current-source loads are made voltage programmable. Assuming the current-source loads are proportional to V_{cntl} with a constant of proportionality of K_{bias}, we have

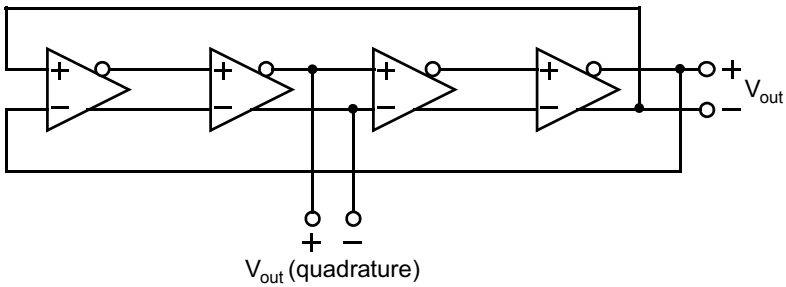


Fig. 19.21 Realizing a ring oscillator using fully differential inverters for improved power-supply rejection.

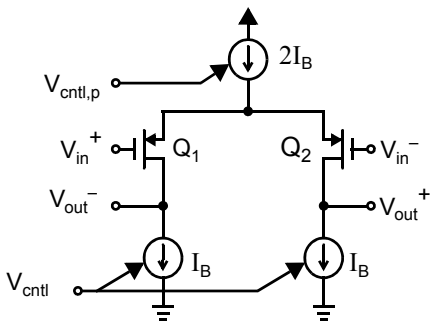


Fig. 19.22 A fully differential inverter with a programmable delay. CMFB circuitry not shown.

$$I_B = K_{bias} V_{cntl} \tag{19.82}$$

The delay of each inverter is proportional to the unity-gain frequency of that inverter. In other words,

$$\tau_{inv} \propto \frac{C_L}{g_m} \tag{19.83}$$

where C_L is the load capacitance of the inverter and g_m is the transconductance of the drive transistors of the inverter. Since $g_m \propto \sqrt{I_B}$, we know that the delay is proportional to $1/\sqrt{I_B} \propto 1/\sqrt{V_{cntl}}$ and therefore, $f_{osc} \propto \sqrt{V_{cntl}}$. Thus, the relationship between the oscillation frequency and the control voltage is not very linear. In a PLL, this nonlinearity implies that K_{osc} is a function of f_{osc} and, therefore, so is the loop bandwidth and settling time. The major reason for this nonlinearity is that at higher frequencies, the voltage changes at the outputs of each stage increase in amplitude because I_B increases.

It is possible to realize a ring-oscillator VCO with good voltage-to-frequency linearity if one ensures that the voltage changes of the delay elements are independent of the oscillation frequency [Kim, 1990; Young, 1992; Reynolds, 1994]. A differential delay stage where this linearization has been realized is shown in Fig. 19.23 [Young, 1992], along with the biasing circuitry that is used to control the voltage swings. Each stage consists of a differential amplifier with p-channel input resistors and resistive loads. The resistive loads realized using n-channel transistors biased in the triode region. Their resistance is adjusted by the bias circuit so that when all of I_b is flowing through them, they will have V_{ref} across them, where V_{ref} is obtained from a temperature-independent voltage reference. This is achieved by using a replica of the delay cell in the bias circuit. The replica consists of Q_3, Q_4, R_3, and R_4. In the replica, all of I_b flows through Q_3 and R_3, so Q_4 and R_4 could have been eliminated from the circuit—they have been included in the figure for clarity only. Recalling that R_3 is actually realized by an n-channel transistor, it is seen that the negative feedback loop including the opamp will cause the voltage across R_3 to be equal to V_{ref} when I_b is going through it. Since the output of the opamp is also used to control the resistance of all the triode-region resistances used for the delay cells of the ring oscillator, they will also have

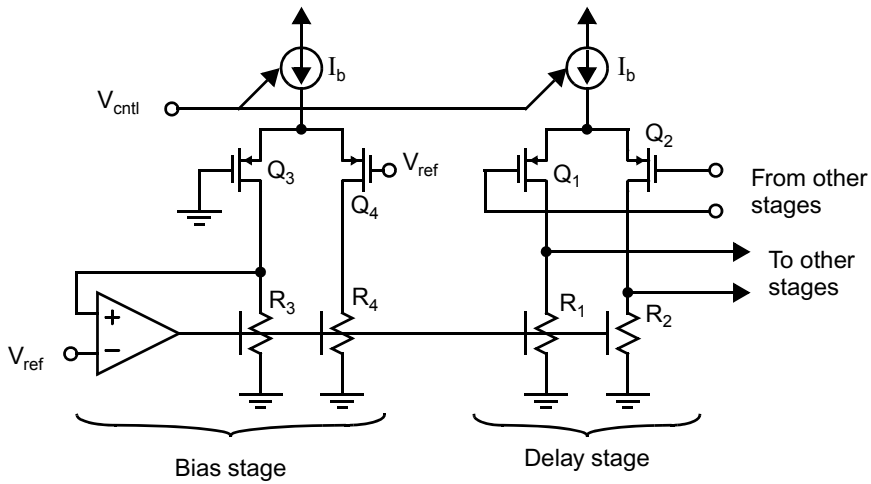


Fig. 19.23 A differential delay cell having resistive loads consisting of n-channel transistors.

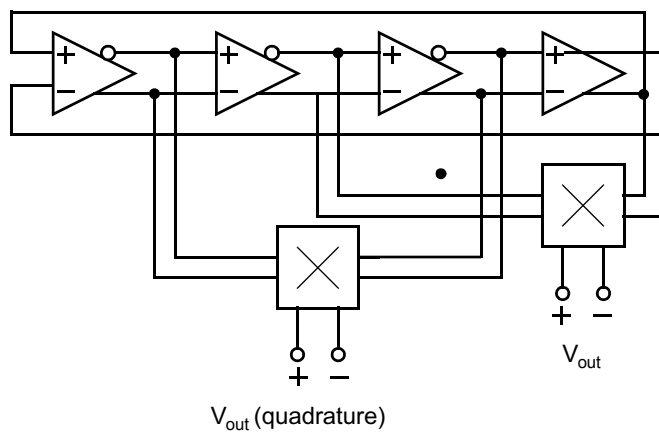


Fig. 19.24 Modifying the ring oscillator of Fig. 19.21 to achieve a doubling of frequency.

V_{ref} across them when I_b flows through them. Next, if the VCO frequency changes because of a change in V_{cntl}, and therefore in I_b, the bias loop will change the resistance of the delay-stage loads so the maximum voltages across them will still equal V_{ref}. Assuming the capacitive loads of each stage are constant (a good assumption), the delay of each stage will be inversely proportional to I_b. This makes the frequency of oscillation proportional to I_b.

In an interesting modification of the ring oscillator of Fig. 19.21, two multipliers have been added in Fig. 19.24. This modification achieves a doubling in the output frequency while still realizing two quadrature outputs [Buchwald, 1992]. Indeed, if quadrature outputs are not required, it is possible to connect the two multiplier outputs to an additional third multiplier. This achieves another doubling of frequency, but there is now only a single output. Another interesting variation of a ring oscillator is described in [Razavi, 1994], where the outputs of

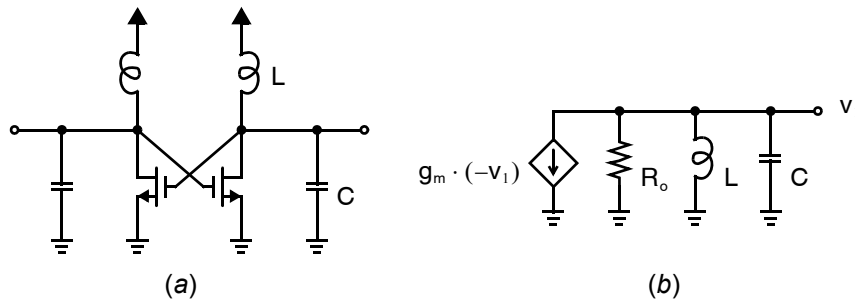


Fig. 19.25 (*a*) A simple CMOS LC-oscillator and (*b*) its differential small-signal half-circuit.

a three-stage oscillator are combined in a novel manner to obtain a single output with a period equal to only two inverter delays.

19.4.2 LC Oscillators

LC oscillators are an example of tuned oscillators. Whereas ring oscillators use active amplifier stages to provide the 180° phase shift required for loop instability, tuned oscillators insert a tuned (resonant) circuit into a feedback loop to provide the phase shift. In the case of an LC oscillator, the resonance is provided by a parallel LC combination.

A simple LC oscillator is shown in Fig. 19.25(*a*) and its small-signal differential half-circuit is in Fig. 19.25(*b*) where R_0 models the output resistance of the transistors as well as any losses associated with the inductor and/or capacitor. Note the similarity with the one-stage ring shown in Fig. 19.19(*b*). The parallel RC network in the one-stage ring provides only 90° phase shift making it stable; the second-order resonant LC tank in Fig. 19.25 provides the additional phase shift required for oscillation. The inductors provide a dc bias voltage to the transistor gates and drains ensuring they are in saturation. The frequency of oscillation is the resonant frequency of the tank.

$$f_0 = \frac{1}{2\pi\sqrt{LC}} \tag{19.84}$$

At resonance, the LC tank has infinite impedance. At that single frequency, the circuit reduces to only R_o in parallel with the transconductor, equivalent to $(R_0||-1/g_m)$. If the negative resistance is larger than R_0, upon startup the amplitude of oscillations at f_o increases exponentially. If, however, the resistance R_0 is larger than $(1/g_m)$ the response decays and the circuit will not oscillate. Hence, to ensure oscillation, the transistors are sized and biased to ensure

$$g_m > \frac{1}{R_0} \tag{19.85}$$

This small-signal analysis is applicable at startup and is therefore useful in finding conditions for oscillation, but the large swings present after oscillations start render it inaccurate (as it predicts the oscillation amplitude grows without bound).

The implementation of the passive tank circuit is critical to the design of an integrated LC oscillator. The capacitors will generally be a combination of parasitic capacitances, varactors to enable tuning of f_0, and perhaps

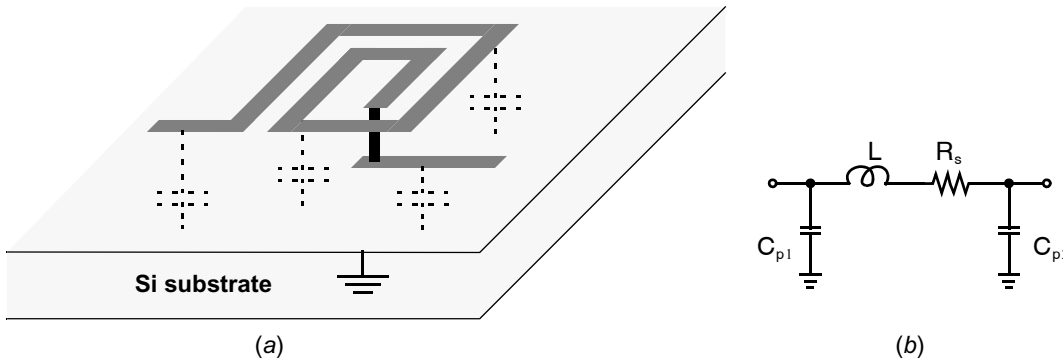


Fig. 19.26 (*a*) A planar integrated inductor may be fabricated on an integrated circuit using the metal intended to interconnect transistors by laying spiral traces above the silicon substrate. (*b*) Its series resistance, R_s, and parasitic capacitances to ground, C_{p1} and C_{p2}, should be included in any circuit simulations involving the inductor.

metal-metal capacitors in series with switches for coarse tuning of f_0. CMOS inductors are formed with spirals of the metal used for interconnect, as pictured in Fig. 19.26(*a*). They are restricted to inductance values of up to a few nH for a spiral several hundred μm per side—considerably smaller than what is possible with discrete inductors due to their smaller physical size and the constraints of modern CMOS fabrication process which preclude the use of magnetic cores. Electrically, inductors behave similar to the model in Fig. 19.26(*b*). In addition to the desired inductance, L, parasitic capacitances to ground and resistive losses must be considered.

Several considerations may influence the designer's choice of whether to use a ring or LC oscillator. The operating frequency of a ring oscillator depends upon the frequency response of the loop (19.81). As a result, it may be tuned over a very wide range of frequencies, up to a decade or more, by changing the biasing of the transistors. By contrast, a LC oscillator's resonant frequency (19.84) can only be influenced by changing the value of the passive components in the tank making it difficult to realize a tuning range of 2:1 or more. However, the insensitivity of an LC oscillator to changes in transistor biasing also means that its oscillations will be more stable (hence, have less jitter) in the presence of temperature and supply voltage variations than those of a ring oscillator. At frequencies within roughly a decade of the maximum f_T of the MOS devices, the power consumption of a ring oscillator can become very high since the individual stages in the ring are required to operate with very high bandwidth. Hence, LC oscillators are generally preferred a frequencies approaching the speed limits of a given CMOS technology. However, at low frequencies the required value of the passive inductor is relatively large so the metal spiral requires a large footprint. As a result, LC-oscillators tend to be larger in circuit area than ring oscillators, a fact which almost completely precludes their use around 1-GHz and below. Finally, ring oscillators generally start oscillating more quickly upon power-up than their LC counterparts. In summary, LC oscillators are preferred when high operating frequency or high spectral purity are required; otherwise ring oscillators offer advantages in area and tuning range.

19.4.3 Phase Noise of Oscillators

Phase noise in oscillators is a fundamental fact of life. Since a lossless oscillator does not exist (that would imply perpetual motion), every oscillator requires some active circuitry to sustain the oscillations and the active circuitry must introduce noise. The specific mechanisms by which device noise becomes phase noise can be subtle and are only recently becoming well understood. However, the phenomenological facts of phase noise are well established, and are sufficient to permit a basic understanding and design of integrated circuit oscillators and PLLs.

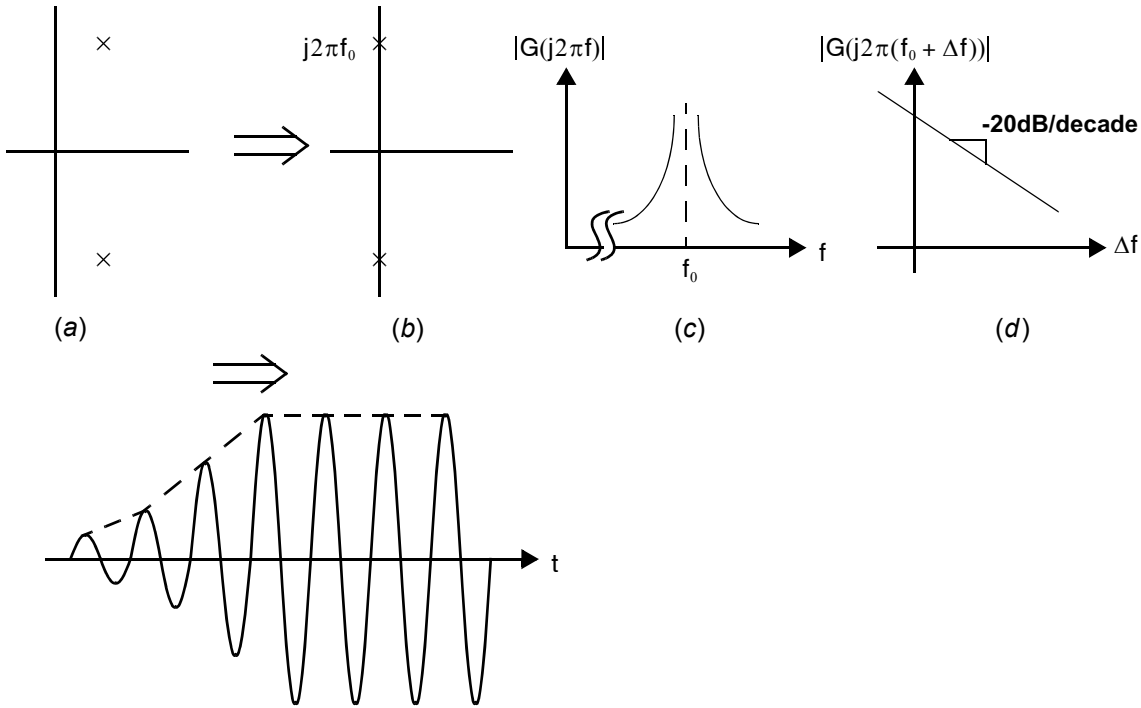


Fig. 19.27 Unstable poles of a closed-loop amplifier: (*a*) poles in the right-half plane give rise to oscillations with an exponentially increasing amplitude; (*b*) as the oscillation amplitude saturates, loop gain is reduced and the poles effectively move to the real axis; (*c*) the frequency response around the pole frequency; (*d*) on a logarithmic scale.

The phase noise of integrated circuit oscillators may be characterized as follows.

$$S_{\phi,\,osc}(f) = h_0 + \frac{h_2}{f^2} + \frac{h_3}{f^3} \tag{19.86}$$

Although a rigorous treatment to derive the three terms in (19.86) is very involved, their existence can be qualitatively understood by simple arguments. The small signal analysis of a ring or LC oscillator will generally reveal that it has poles in the right half plane, as in Fig. 19.27(*a*), so its impulse response is oscillation with an exponentially-growing envelope. The amplitude of these oscillations eventually saturates. Although saturation is a nonlinear phenomena, the result is similar to what occurs in a linear system with poles precisely on the imaginary axis as shown in Fig. 19.27(*b*): The oscillations are sustained with constant amplitude. The frequency response of such a linear model $G(j2\pi(f_0 + \Delta f))$ for small offset frequencies Δf will be roughly constant except for the pole term.

$$G(j2\pi(f_0 + \Delta f)) = \left(\frac{1}{1 - j(f_0 + \Delta f)/jf_0}\right) \cdot (\ldots) = \left(\frac{-f_0}{\Delta f}\right) \cdot (\ldots) \tag{19.87}$$

Noise sources within the oscillator are shaped by $|G(j2\pi(f_0 + \Delta f))|^2$, which from (19.87) is proportional to $1/\Delta f^2$. The discussion so far explains $1/\Delta f^2$-shaping of noise sources in $\mathcal{L}(\Delta f)$, but in practice such shaping is also observed in the phase noise spectrum, as predicted by (19.76).

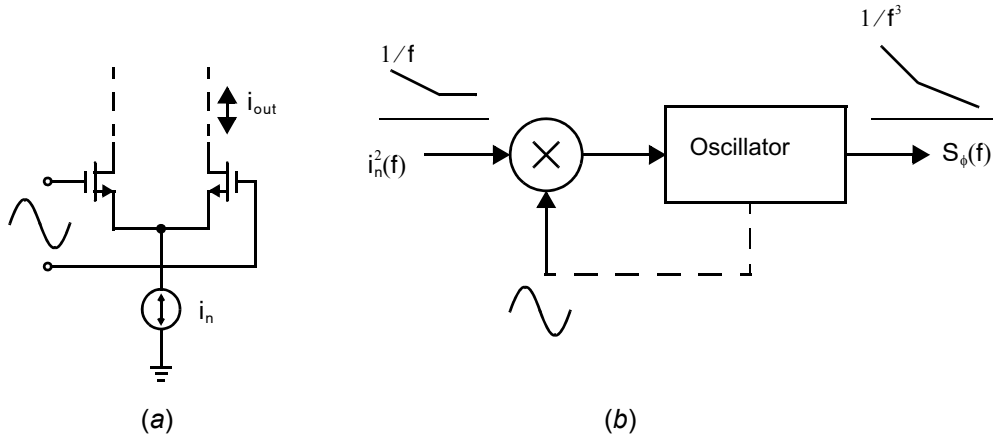


Fig. 19.28 Up-conversion of noise sources by switching in an oscillator.

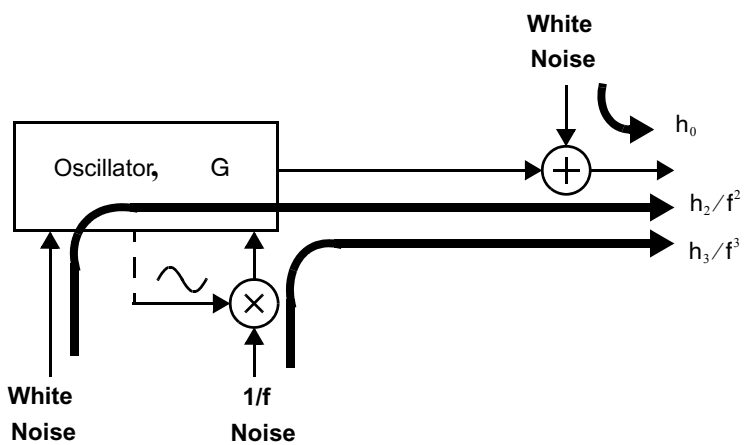


Fig. 19.29 Major sources of phase noise in integrated circuit oscillations.

Hence, white noise introduced within the oscillator's closed feedback loop is shaped by the oscillator's response giving rise to the h_2/f^2 term in (19.86). Some sources of noise are switched by the oscillations resulting in an up-conversion of their noise spectral density to the carrier frequency before being shaped by the oscillator response. A classic example is illustrated in Fig. 19.28 where noise is introduced by the tail current transistor in a differential pair with oscillating inputs. Flicker noise up-converted in this way, and then spectrally shaped by the closed-loop oscillator response gives rise to the h_3/f^3 term in (19.86). Finally, it was shown in Example 19.9 that additive white noise superimposed on a sinusoid gives rise to white phase noise. This is the situation when white noise is introduced immediately following an oscillator core, for example by a buffer, and the resulting phase noise is modeled by h_0 in (19.86). The superposition of these three effects is diagrammed in Fig. 19.29.

A logarithmic plot of phase noise based on (19.86) is sketched in Fig. 19.30. The corner frequency f_c marks the transition between the $1/f^2$-dominated region and the $1/f^3$-dominated region and is given by the frequency at

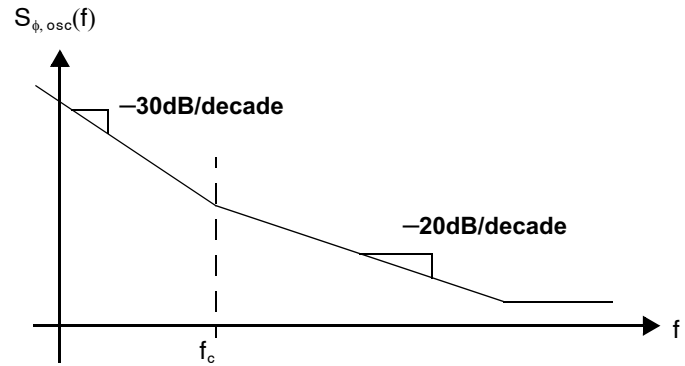


Fig. 19.30 Typical phase noise of an integrated circuit oscillator.

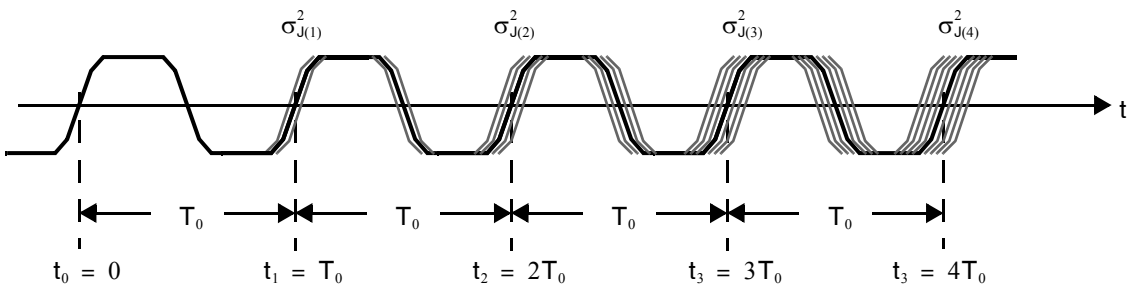


Fig. 19.31 The accumulation of jitter over multiple cycles.

which these two terms contribute equally to S_ϕ.

$$f_c = \frac{h_3}{h_2} \tag{19.88}$$

A potentially troubling aspect of the phase noise model in (19.86) is that the close-in phase noise terms $1/f^2$ and $1/f^3$ cause the integrals in (19.60), (19.64), (19.66), and (19.68) to diverge. The absolute jitter (19.60) of an oscillator does, in fact, diverge—another reason that it is a mostly analytical tool. However, the period, P-cycle, and adjacent period jitter of an oscillator may be measured in a lab and are finite, in contrast to the model's predictions. Hence, the simple phase noise model of (19.86) is considered inaccurate at very small offset frequencies, f. The problem is that defining phase noise as the noise spectral density of ϕ_k is strictly valid only if phase is a stationary random process. In fact, it is not [Gardner, 2005]. When observed over long periods of time (on the order of seconds) this non-stationarity causes analytical problems, so phase noise is not usually considered below a few Hertz. When using (19.64), (19.66) and (19.68) to estimate jitter a lower integration limit of a few Hertz is therefore applied [Liu, 2004]. Above this limit the phase noise model of (19.86) is reliable and can be integrated to accurately estimate jitter.

To understand the unboundedness of the absolute jitter of an oscillator, consider how the P-cycle jitter evolves for increasing P. As shown in Fig. 19.31, as P increases, the variance of the oscillator's P-th consecutive zero crossing time increases. Qualitatively, this is because disturbances in the state of the oscillator due to internal noise persist, circulating around its closed feedback loop. Hence, the jitter becomes a superposition of noise contributions for

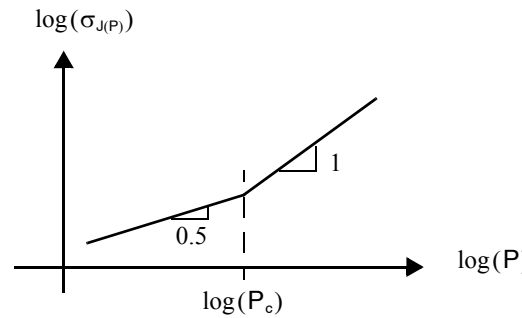


Fig. 19.32 The evolution of P-cycle jitter over time.

all past time and grows without bound.[16] Looking at the expression for P-cycle jitter in (19.66), with small P the $\sin^2(\pi f P T_0)$ term in the integrand filters out low-frequency phase noise, $S_\phi(f)$. As P increases, more low-frequency phase noise is included in the integral and the resulting P-cycle jitter therefore grows.

A logarithmic plot of $\sigma_{J(P)}$ versus P is sketched in Fig. 19.32 [Liu, 2004]. The h_2/f^2 component of phase noise causes P-cycle jitter to increase in proportion to $\sqrt{P}$. Including only this term in (19.66), it may be shown that

$$\sigma_{J(P)} = \sqrt{h_2 T_0^3} \cdot \sqrt{P} \tag{19.89}$$

Once N becomes large enough for frequencies below f_c to have a significant effect, the h_3/f^3 term causes P-cycle jitter to grow approximately in direct proportion to P with the transition between the two behaviors at

$$P_c \cong \frac{1}{25 T_0 f_c}. \tag{19.90}$$

Equations (19.89) and (19.90) allow for estimates of the phase-noise profile of an integrated oscillator, Fig. 19.30, from the time-domain measurements in Fig. 19.32.

EXAMPLE 19.18

The P-cycle jitter of a VCO operating at 1.5 GHz is observed for different P to obtain a plot like Fig. 19.32 with $P_c = 30$. The P-cycle jitter over 10 cycles ($P = 10$) is 2.0 ps rms. Estimate the phase noise of the VCO using the model in (19.86) but neglecting the white-phase noise term h_0, which is only important at very high frequencies.

Solution

Substituting $P = 10$, $\sigma_{J(10)} = 2$ ps, and $T_0 = (1/1.5 \cdot 10^9)$ s into (19.89) yields an estimate of the phase noise model parameter $h_2 \cong 1350 \text{ rad}^2 \cdot \text{Hz}$. Using (19.90) with $P_c = 30$ yields $f_c \cong 2$ MHz. Finally, substitution into (19.88) yields $h_3 \cong 2.7 \cdot 10^9 \text{ rad}^2 \cdot \text{Hz}^2$. Hence, the phase noise of the VCO may be modeled as follows:

$$S_{\phi,\,osc}(f) = \frac{1350 \text{ rad}^2 \cdot \text{Hz}}{f^2} + \frac{2.7 \cdot 10^9 \text{ rad}^2 \cdot \text{Hz}^2}{f^3} \tag{19.91}$$

16. This is analogous to a random-walk process: a person beginning at a fixed point takes a step (randomly) in any direction. Over time, their mean position remains equal to their starting point since their direction of travel is random, but the variance of their position continues to grow as they may (with finite probability) wind up very far from the their starting point.

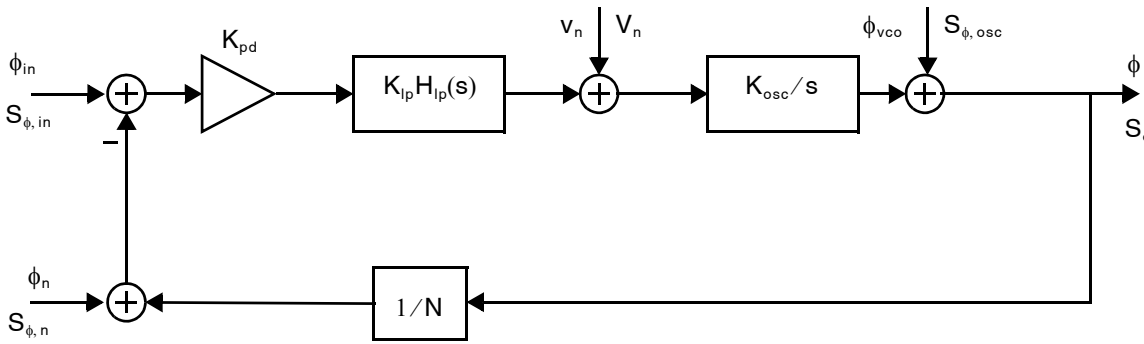


Fig. 19.33 Linear model of an integer-N PLL showing noise sources.

19.5 JITTER AND PHASE NOISE IN PLLS

There are several sources of jitter in a PLL. Notably:

- Jitter on the input reference, ϕ_{in}
- Jitter inherent in the VCO
- Noise generated by circuitry in the loop filter
- Noise generated by circuitry in the divider

Since the jitter in any practical PLL is relatively small, its propagation through and within the loop can be analyzed using a linear small-signal model. The noise sources listed above are introduced at various points in the loop, as shown in the linear model block diagram in Fig. 19.33. In order for that diagram to be general, all noise in the phase detector and loop filter must be modeled by the single output-referred noise voltage, v_n, and the additional jitter introduced by the divider is modeled by the additive term ϕ_n. Also labeled in Fig. 19.33 are the power spectral densities of all the random signals: the noise spectral density V_n, and the phase noise spectra $S_{\phi,\,in}$, $S_{\phi,\,n}$, $S_{\phi,\,osc}$, and S_ϕ. Each noise source's contribution to the PLL output phase noise is filtered by a different response.

19.5.1 Input Phase Noise and Divider Phase Noise

Both the input reference phase noise and divider phase noise appear at the input and are therefore filtered by the PLL's jitter transfer function,

$$H(s) = \frac{K_{pd}K_{lp}K_{osc}H_{lp}(s)/s}{1 + L(s)} \tag{19.92}$$

where $L(s)$ is the PLL's open loop gain given by (19.30). For the special case of a 2nd order PLL, (19.92) becomes (19.35).

Since the loop filter $H_{lp}(s)$ is lowpass, $H(s)$ will also be lowpass. A dc gain of N arises because the output frequency is N-times higher than the input; hence, a phase variation of $\Delta\phi_{in}$ rad at the input that is perfectly tracked at the output becomes a $N\Delta\phi_{in}$ rad phase change with respect to the higher output frequency. Hence, any slow variations in the timing of the input clock will be faithfully reproduced in the output clock—exactly what one would expect from a phase-locked system. However, very fast variations in the input clock phase, those that lie beyond the PLL's closed-loop bandwidth, are not tracked. Rather, they are attenuated by the lowpass $|H(f)|^2$.

19.5.2 VCO Phase Noise

The VCO phase noise is filtered by a highpass response straightforwardly obtained from the signal flow graph of Fig. 19.33.

$$H_{osc}(s) = \frac{1}{1 + L(s)} \tag{19.93}$$

At high frequency, $|L(f)|^2 \approx 0$ and VCO phase noise appears unaltered at the output, $H_{osc}(s) \approx 1$. However, at low frequency the loop gain is large and VCO phase noise is attenuated. At first sight, it may appear strange that the VCO's inherent phase noise somehow disappears when enclosed within a PLL.

The PLL senses variations in the VCO's output phase and adjusts the input V_{cntl} to counter those variations in an effort to keep the output phase-locked to the input. Hence, slow natural variations in the VCO's phase are completely canceled by the loop. On the other hand, the loop is unable to respond to very fast variations, which appear identically in the high-frequency portion of $S_{\phi,osc}(f)$ and $S_{\phi}(f)$.

EXAMPLE 19.19

A VCO with a phase noise given by $S_{\phi,vco}(f) = h_2/f^2$ is placed within an ideal second-order PLL. What is the resulting phase noise at the output?

Solution

For a second-order PLL, the open-loop gain $L(s)$ is specified in (19.33). Substitution into (19.93) yields the transfer function from VCO phase noise to the PLL output.

$$H_{osc}(s) = \frac{s^2}{s^2 + \frac{\omega_{pll}}{Q}s + \omega_{pll}^2} \tag{19.94}$$

The output phase noise is given by,

$$\begin{aligned} S_{\phi}(f) &= S_{\phi,osc}(f)|H_{osc}(f)|^2 \\ &= \frac{h_2}{f^2} \cdot \frac{f^4}{\left(f^2 + \frac{\omega_{pll}}{2\pi Q}f + \frac{\omega_{pll}^2}{4\pi^2}\right)^2} \\ &= \frac{h_2 f^2}{\left(f^2 + \frac{\omega_{pll}}{2\pi Q}f + \frac{\omega_{pll}^2}{4\pi^2}\right)^2} \end{aligned} \tag{19.95}$$

This is a bandpass spectrum with a peak around ω_{pll}. If (19.95) is substituted into (19.60) with the upper limit of integration approximated as infinite ($1/2T_0 \approx \infty$) a very simple expression is obtained for absolute jitter.

$$\sigma_{\tau}^2 = \frac{h_2 Q T_0^2}{2\omega_{pll}} \tag{19.96}$$

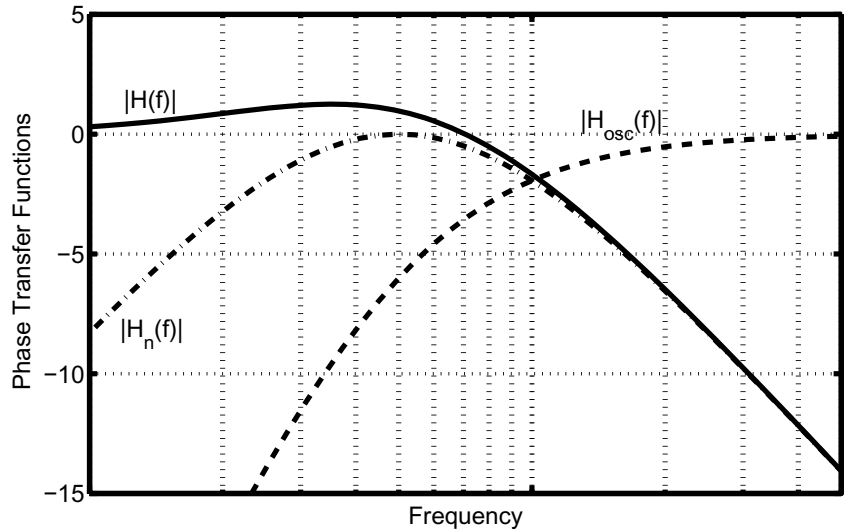


Fig. 19.34 Normalized transfer functions from noise sources within a PLL to the output phase noise.

This expression reveals that the PLL loop bandwidth (approximately equal to ω_{pll}/Q) should be maximized in order to minimize the contribution of VCO phase noise to a PLL's output jitter.

19.5.3 Loop Filter Noise

Looking at Fig. 19.33, the loop filter noise V_n is filtered by the following response:

$$H_n(s) = \frac{K_{osc}/s}{1 + L(s)} \tag{19.97}$$

At high frequencies due to the lowpass H_{lp}, $H_n \approx K_{osc}/s \approx 0$. At low frequencies, $H_n(s) \approx N/K_{pd}K_{lp}H_{lp}(s)$. Typically, the combination of the phase detector and loop filter $K_{pd}K_{lp}H_{lp}(s)$ has very high gain at dc to ensure precise phase locking of the input and output in steady-state, so H_n is also very small at dc. For example, in the case of an ideal charge-pump loop filter H_{lp} has a pole at dc, so that $H_n(0) = 0$. Hence, noise introduced in the loop filter appears bandpass-filtered in the output phase noise.

Key Point: *PLL loop bandwidth sets the cutoff frequency below which VCO phase noise is attenuated and above which input reference (and divider) phase noise is attenuated.*

Typical PLL responses $|H(f)|$, $|H_{osc}(f)|$, and $|H_n(f)|$ are sketched in Fig. 19.34. The PLL loop bandwidth sets the cutoff frequency below which VCO phase noise is attenuated and above which input reference (and divider) phase noise is attenuated.

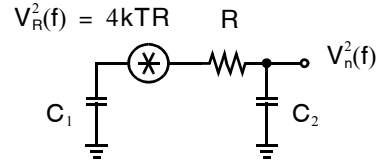


Fig. 19.35 Equivalent circuit for the noise analysis of a charge-pump loop filter.

EXAMPLE 19.20

Consider the charge-pump PLL designed in Example 19.8 and consider its operation with $N = 75$ using the VCO in Example 19.18. Assume the 20-MHz PLL input (after the $M = 2$ divider) has a phase noise modeled by (19.86) with $h_0 = 10^{-14}$ rad^2/Hz, $h_2 = 10^{-9}$ rad$^2 \cdot$ Hz, and $h_3 \approx 0$. What is the phase noise of the PLL output due to the reference input, loop filter resistor, and VCO?

Solution

The contribution of the reference input to the PLL output phase noise is given by

$$S_{\phi,\,in}(f) = \left(8 \cdot 10^{-16} + \frac{7 \cdot 10^{-11}}{f^2}\right) \text{ rad}^2/\text{Hz}$$

filtered by $|H(f)|^2$. If C_2 is neglected, $H(s)$ is given by (19.35) with the values of ω_{pll} and $\omega_{z'}$ being those chosen in Example 19.8: $2\pi \cdot 800$ kHz and $2\pi \cdot 400$ kHz respectively.

The loop filter resistor R results in thermal noise at the control voltage node, V_{cntl}. Its noise spectral density may be determined by a noise circuit analysis of Fig. 19.35 resulting in

$$V_n^2(f) = 4kTR\left(\frac{C_{eq}}{C_2}\right)^2\left(\frac{1}{1 + f^2R^2C_{eq}^2}\right) \tag{19.98}$$

where $C_{eq} = C_1C_2/(C_1 + C_2)$. Maintaining the 2nd-order model, this noise is then bandpass filtered by

$$H_n(s) = \frac{K_{osc}s}{s^2 + \frac{\omega_{pll}}{Q}s + \omega_{pll}^2} \tag{19.99}$$

where $Q = 0.5$ from Example 19.8.

Finally, from Example 19.18 the VCO phase noise modeled by (19.91) is filtered by $|H_{osc}(f)|^2$ with H_{osc} as in (19.94).

The total output phase noise is the superposition of all contributors.

$$S_\phi(f) = S_{\phi,\,in}(f)|H(f)|^2 + V_n^2(f)|H_n(f)|^2 + S_{\phi,\,osc}(f)|H_{osc}(f)|^2 \tag{19.100}$$

A plot of all three phase noise contributors and their sum is shown in Fig. 19.36. Note that at low frequencies, the total output phase noise is equal to the input reference phase noise because the other noise contributors are filtered out by the PLL. Similarly, at high frequencies only the VCO phase noise contributes. In this example, thermal noise in the loop filter plays an insignificant role in the total output phase noise—a typical result. The jitter of the PLL output will be dominated by the high-frequency VCO noise contributions which span a much wider range of frequencies than the input-dominated portion of the phase noise spectrum and, hence, contribute much more to the integral in (19.60).

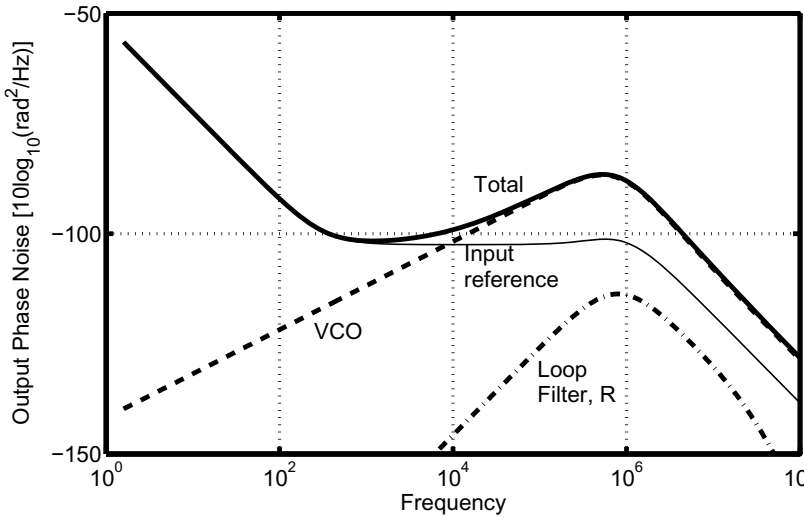


Fig. 19.36 Phase noise contributions of the input reference, loop filter resistor, and VCO to the output of the PLL in Example 19.20.

19.6 KEY POINTS

- The instantaneous frequency deviation of a VCO from its free-running frequency is the derivative of its instantaneous phase. Hence, its instantaneous phase is the integral of the instantaneous frequency deviation. [p. 740]
- Once a PLL is in lock, its dynamic response to input phase and frequency changes can be well approximated by a linear model, as long as these changes are slow and small about their operating or bias point. [p. 748]
- The design of a second-order PLL with Q = 0.5 provides good transient settling behavior without excessive spreading of the loop filter time constants. A design with Q = 0.1 offers even better tracking behavior, but generally demands a larger loop-filter time constant for the same loop bandwidth. [p. 750]
- Due to the discrete-time nature of the PLL, its loop bandwidth is limited by stability considerations to at most roughly 1/20th to 1/10th of the reference frequency, a fact not predicted by the continuous-time linear model. [p. 752]
- Jitter is a measure of random variations in the timing phase of a clock signal and phase noise is its frequency-domain representation. They present fundamental limitations on the accuracy of time references. [p. 756]
- The absolute jitter of a clock, τ_k, is a discrete-time sequence giving the displacement of its transition times away from their ideal values. [p. 756]
- Periodic variations in the phase of a clock result in discrete tones in phase noise spectra and are called spurs. [p. 762]
- PLL loop bandwidth sets the cutoff frequency below which VCO phase noise is attenuated and above which input reference (and divider) phase noise is attenuated. [p. 779]

19.7 REFERENCES

A. W. Buchwald and K. W. Martin. "High-Speed Voltage-Controlled Oscillator with Quadrature Outputs," *Electron. Lett.,* Vol. 27, no. 4, pp. 309–310, February 1991.

A. Buchwald, K. Martin, A. Oki, and K. Kobayashi. "A 6-GHz Integrated Phase-Locked Loop Using AlGaAs/GaAs Hetrojunction Bipolar Transistors," *IEEE J. Solid-State Circuits,* Vol. 27, no. 12, pp. 1752–1761, December 1992.

B. Kim, D. Helman, and Paul Gray. "A 30-MHz Hybrid Analog-Digital Clock Recovery Circuit Is 2-μm CMOS, *IEEE J. of Solid-State Circuits,* Vol. 25, no. 6, pp. 1385–1394, December 1990.

N. M. Nguyen and R. G. Meyer. "A 1.8-GHz Monolithic LC Voltage-Controlled Oscillator," *IEEE J. Solid-State Circuits,* Vol. 27, no. 3, pp. 444–450, March 1992.

B. Razavi and J. Sung. "A 6-GHz 60-mW BiCMOS Phase-Locked Loop, *IEEE J. Solid-State Circuits,* Vol. 29, no. 12, pp. 1560–1571, December 1994.

D. Reynolds. "A 320-MHz CMOS Triple 8-bit DAC with On-Chip PLL and Hardware Cursor," *IEEE J. of Solid-State Circuits,* Vol. 29, no. 12, pp. 1545–1551, December 1994.

I. Young, J. Greason, and K. Wong. "A PLL Clock Generator with 5 to 110 MHz of Lock Range for Microprocessors," *IEEE J. of Solid-State Circuits,* Vol. 27, no. 11, pp. 1599–1607, November 1992.

C. Liu and J. A. McNeill. "Jitter in Oscillators with 1/f Noise Sources," *International Symposium on Circuits and Systems*, 2004.

F. Gardner. *Phaselock Techniques*, John Wiley & Sons Inc., 2005.

19.8 PROBLEMS

19.8.1 Section 19.1: Basic Phase-Locked Loop Architecture

19.1 A VCO having $K_{osc} = 2\pi \times 100 \cdot 10^6$ rad/Vs oscillates at a frequency of 1 GHz with $V_{cntl} = 0$. A rectangular pulse appears on V_{cntl} with an amplitude of 1 Volt and a duration of 0.5 ns. What is the resulting phase change at the output of the VCO?

19.2 A VCO has a free-running frequency of 600 MHz with $V_{cntl} = 0$. Its output is applied to a frequency divider with $N = 2$. When a pulse of amplitude 100 mV and duration 1 ns is applied to V_{cntl}, the phase of the 300 MHz clock at the output of the divider is observed to change by 40 degrees. Estimate K_{osc} in units of rad/V.

19.3 Find an expression in terms of ω_0 and K_{osc} for the number of clock cycles required for the output phase of a VCO to change by 2π rad after V_{cntl} is changed from 0 to 1 mV.

19.4 Select the loop filter components R, C_1, and C_2 in a charge-pump PLL so that the maximum capacitance is 30 pF, $C_2 = 0.05C_1$, and to provide a zero at 100 kHz.

19.5 A PLL with $N = 5$ accepts a reference clock input at 20 MHz. The VCO has a free-running frequency of 80 MHz and $K_{osc} = 2\pi \times 40 \cdot 10^6$ rad/Vs. Find the steady-state value of V_{cntl} in lock.

19.6 A PLL with $N = 20$ accepts a reference clock input at 40 MHz and uses a VCO whose free-running frequency varies over the range 700–900 MHz with changes in temperature. Find the value of K_{osc} required to ensure that $-0.5\ \text{V} < V_{cntl} < 0.5\ \text{V}$ when the PLL is in lock.

19.8.2 Section 19.2: Linearized Small-Signal Analysis

19.7 Verify expressions (19.42) and (19.43).

19.8 Produce accurate Bode plots of the open loop response $L(s)$ of a charge-pump PLL for the two cases below, both with and without including C_2. In each case, find the impact of C_2 upon the phase margin of the loop.

a. $Q = 0.5$ and $C_2 = C_1/10$

b. $Q = 0.1$ and $C_2 = C_1/250$

19.9 A PLL is to operate with a reference clock frequency of 50 MHz and an output frequency of 2 GHz. Design a loop filter for the charge-pump phase comparator so that the loop bandwidth is approximately 1/15th of the reference clock frequency and $Q = 0.1$. Assume the VCO has $K_{osc} = 2\pi \times 150 \cdot 10^6$ rad/Vs and the charge-pump current is $I_{ch} = 15\ \mu A$.

19.10 A PLL is to operate with a reference clock frequency of 10 MHz and an output frequency programmable over the range 450 - 550 MHz by changing the divide ratio, N. What value of N results in the worst case loop bandwidth? Assuming $K_{osc} = 2\pi \times 75 \cdot 10^6$ rad/Vs, what range of voltages must be accommodated at V_{cntl}? Design the charge pump loop filter to have a loop bandwidth equal to 1/15th of the reference clock frequency and $Q = 0.5$ when the output frequency is set to 500 MHz and $I_{ch} = 50\ \mu A$. Find the new values of Q when the output frequency is changed to 450 MHz and 550 MHz without changing the loop filter.

19.11 A charge pump PLL must provide a 1 GHz output and a loop bandwidth of $\omega_{3dB} = 1$ MHz while operating from a 20 MHz reference input. With $I_{ch} = 40\ \mu A$ and $K_{osc} = 2\pi \times 20 \cdot 10^6$ rad/Vs, find the values of the loop filter components R, C_1, and C_2 to have $Q = 0.1$. Then, the input reference frequency is changed to 25 MHz. The same output frequency and loop bandwidth are maintained while changing only the values of N and I_{ch}. Find the corresponding new values of N, I_{ch}, and Q.

19.12 An additional pole may be introduced into a charge pump PLL by modifying the loop filter as shown in Fig. P19.12. Find expressions for $K_{lp}H_{lp}(s)$ and all of its poles and zeros.

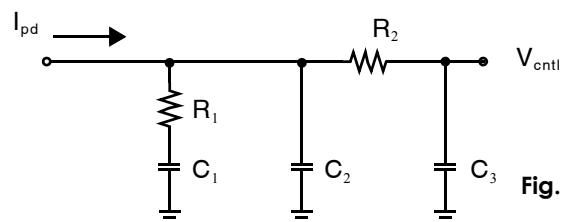


Fig. P19.12

19.13 Another popular charge pump PLL architecture is shown in Fig. P19.13. In this architecture, the VCO accepts two inputs: V_{cntl1} and V_{cntl2}. Each influences the instantaneous frequency via a different gain:

$$\omega = K_{osc1}V_{cntl1} + K_{osc2}V_{cntl2}$$

Find expressions for the open loop gain, $L(s)$, and closed loop jitter transfer function, $H(s)$, of this PLL in terms of the loop parameters, I_{ch1}, I_{ch1}, R_1, C_1, etc.

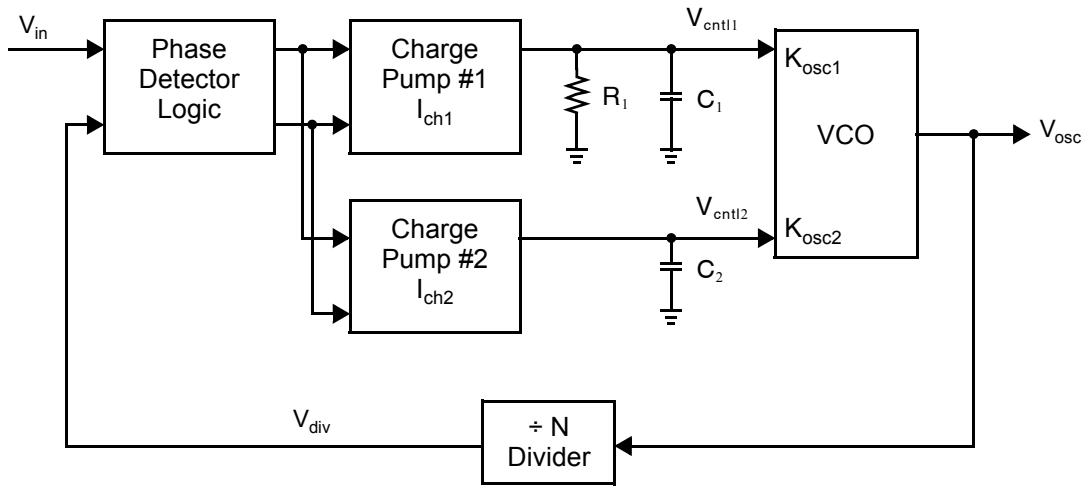


Fig. P19.13

19.8.3 Section 19.3: Jitter and Phase Noise

19.14 A clock waveform at a nominal frequency of 125 MHz is observed to have zero-crossings at the following times:

$$t_k = 0, 8.1 \text{ ns}, 16.0 \text{ ns}, 22.9 \text{ ns}, 31.4 \text{ ns}, 40.2 \text{ ns}, 41.0 \text{ ns}, 49.0 \text{ ns}, 56.8 \text{ ns}, 64.1 \text{ ns}$$

Find the absolute jitter sequence, τ_k, the period sequence, T_k, the period jitter sequence, J_k, and the 3-cycle jitter sequence, $J(3)_k$.

19.15 A clock voltage waveform at frequency f has amplitude A and rise-fall times of T_{rf}, as illustrated in Fig. P19.15. Superimposed is gaussian voltage noise with variance σ_n^2. Find expressions for the variance of the absolute jitter in seconds, σ_τ^2, and in radians, σ_ϕ^2.

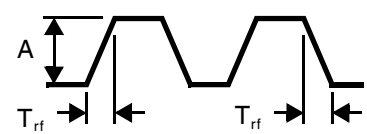


Fig. P19.15

19.16 Consider the common source amplifier of Fig. 19.14 where transistor Q_1 is biased in saturation by dc source V_{bias} and v_s is a noiseless small-signal sinusoid of amplitude $|V_s|$. Find the absolute jitter in seconds, σ_τ^2, and in radians, σ_ϕ^2, that results from power supply noise having a variance σ_n^2 V^2. You may neglect all other noise sources and parasitic capacitances.

19.8.4 Section 19.4: Electronic Oscillators

19.17 A ring oscillator made from n identical stages modeled as in Fig. 19.19(*b*) has an open loop response $L(s)$ given by (19.81). Find an expression for the frequency at which L exhibits $180°$ phase shift in terms of n, G_m, R_o, and C.

19.18 For a ring oscillator with $n = 3$ identical first-order stages modeled as in Fig. 19.19(*b*), find an expression for the minimum dc gain per stage required to ensure that the ring is unstable.

19.19 An 2 GHz LC oscillator is to be designed as in Fig. 19.25 using a 1 nH inductor and with $R_o = 1\ \text{k}\Omega$. Find the required total tank capacitance, C. Estimate the required power consumption, assuming the supply voltage is 1.5 V and the transistors are biased so that $V_{eff} = 350$ mV.

19.20 The varactor in an LC VCO can change its capacitance by at most $\pm 15\%$ with changes in V_{cntl}. What is the resulting percentage change in the VCO's oscillation frequency?

19.21 The capacitance of a varactor changes linearly with its control voltage by ν Ferads/V. Find an expression for the gain of an LC VCO, K_{osc}, using such a varactor in terms of ν and L.

19.22 A LC VCO is to be designed using a 500 pH inductor and a varactor to cover the range of frequencies 2.8 GHz to 3.2 GHz with $-1 \text{ V} < V_{cntl} < 1 \text{ V}$. What range of capacitance values must C cover? Express this range as a percentage of the nominal capacitance at C. Assuming that C varies linearly with V_{cntl} over the range $-1 \text{ V} < V_{cntl} < 1 \text{ V}$, by what percentage does K_{osc} change going from 2.8 GHz to 3.2 GHz?

19.8.5 Section 19.5: Jitter and Phase Noise in PLLs

19.23 An input clock at 2.5 GHz is known to have a significant spur at 100 MHz offset. We wish to design a charge pump PLL to suppress that spur by at least 60 dB. Assuming the filter has $Q = 0.5$, what should be the bandwidth of the PLL?

19.24 For a particular 1.5 GHz PLL designed to have $Q = 0.5$ and a loop bandwidth of 500 kHz, it is required that the VCO's inherent phase noise contribute only 10 ps rms to the absolute jitter at the PLL output. Find a specification on h_2 assuming the VCO's phase noise is adequately modeled by $S_{\phi,vco}(f) = h_2/f^2$.

19.25 A VCO's phase noise is $S_{\phi,vco}(f) = h_2/f^2$ with $h_2 = 500 \text{ rad}^2 \cdot \text{Hz}$. Find the loop bandwidth of a charge pump PLL with $Q = 0.1$ so that the absolute jitter contributed by the VCO to the output of the PLL is 0.01 radians rms.

19.26 Find the peak value of $|H_n(f)|$, the frequency response from loop filter noise to PLL output phase noise in (19.99), in terms of the PLL parameters, K_{osc}, Q, etc.

19.27 Assume that power supply noise introduces white noise onto V_{cntl} that may be modeled as in Fig. 19.33 with $V_n^2(f) = 10^{-12} \text{ V}^2/\text{Hz}$. Find the maximum VCO gain, K_{osc}, so that with $Q = 0.5$ the contribution of this power supply noise to the PLL's output phase noise never exceeds $10^{-11} \text{ rad}^2/\text{Hz}$.

19.28 Find an expression for the contribution of resistor R_1 to the output phase noise for the PLL depicted in Fig. P19.13.

Index

1/f noise, *see* noise, 1/f
1/f noise tangent, *see* noise, 1/f noise tangent
1's complement, 612–614
2's complement, 606, 612–614, 717

A

A/D converter:
  algorithmic, 647, 662–665, 666, 667
  charge-redistribution, 653–659
  cyclic, *see* A/D converter, algorithmic
  dual slope, 646–648
  flash, 647, 673–678, 680, 682–685
  folding, 647, 680, 683–687
  ideal, 608–609, 611, 615–618
  integrating, 646–650
  interpolating, 647, 675, 680–683, 685–686
  Nyquist-rate, 606, 646–695, 711
  oversampling, 537, 594, 606, 611, 618, 646, 647, 696–737
  pipelined, 588, 617, 647, 664, 665–673, 678, 679
  quad-slope, 648
  single-slope, 648
  successive-approximation, 647, 650–662, 665
  time-interleaved, 647, 687–689
  two-step, 647, 677–680, 683, 686
acceptor, 2
accumulated channel, 16–17
accuracy:
  absolute, 615, 618–619, 679, 680
  relative, 615, 616, 618, 619
active load, 120, 122, 124, 126, 136, 138, 196, 224, 243, 244, 264, 283, 395
active-RC, *see* filter, active-RC
active region:
  bipolar operation, 333–335, 337–338, 462, 464
  MOSFET layout, 86–88, 90–93, 97–98, 484
  MOSFET operation, 20–34, 41, 118, 120, 129, 130, 132, 165–167, 194, 262, 379–381, 425, 493–499, 509
adaptive filter, 523–524
algorithmic A/D, *see* A/D converter, algorithmic
aliasing, 469, 519, 537, 540, 543–544, 553–557, 606, 696, 702, 711, 717, 721, 760
amorphous silicon, 15, 82
annealing, 76–77, 84, 96, 347–348
antenna rules, 94–95
anti-aliasing, *see* aliasing
aperture jitter 445, 617, 620
aperture uncertainty, *see* aperture jitter
arsenic 2, 76, 77, 79, 82, 83

B

bandgap, 306, 310–320, 321
beat test, 445, 446
bias circuit, 243, 260, 299, 302–304, 307–310, 767, 769
BiCMOS, 86, 290, 331, 346–348, 433–436, 452, 453, 456–460, 506–509, 557, 626
bilinear transform, 550–552, 573–574
binary search, 651–652, 662
binary-scaled D/A, *see* D/A converter, binary-scaled
bipolar-junction transistors (BJTs), 7, 13, 14, 20, 38, 42, 49, 86, 95–96, 290, 311, 315, 331–346, 348–349, 356–359, 378, 380
bipolar process, 310, 346–349
biquadratic filter, *see* filter, second-order
body effect, 20, 24, 26, 30, 51–52, 64–66, 68, 81, 120–128, 130, 179, 188, 190, 244, 263, 308, 353, 381, 447, 559
Boltzmann's constant, 4, 38, 311, 378
boron, 2, 74, 76, 77, 79, 81, 83, 347
BSIM model, 52
bubble error, 674, 676–677, 686
built-in junction potential, 3, 4, 50–52, 56, 341, 345–346
buried n well, 85
buried collector, 332
butting contact, 90

C

calibration, 522, 593, 638–640, 648, 659–661, 702, 728–729
capacitor:
  metal-metal, 59–60

MOS, 58–59
pn junction, 6–13, 58
cascade:
current mirror, 129–131, 309–310
folded, 131–133, 181, 186–187, 252, 268–269, 271–274, 279–280, 283–284, 286–288, 291, 293
gain stage, 131–133, 138, 181–182, 187, 198, 271
opamp, 242, 252, 268–274, 277, 279–280, 283–284, 286–288, 291, 293
telescopic, 131–132, 181, 186–187
channel-length modulation, 21–23
charge injection 18, 292, 413, 418–424, 447, 450, 451, 454, 559, 585–588, 592, 638, 640
charge-redistribution:
A/D, *see* A/D converter, charge-redistribution
D/A, *see* D/A converter, charge-redistribution
chemical vapor deposition (CVD), 78, 80, 84–85
chopper stabilization, 252
clock synthesis, *see* frequency synthesis
CMOS pair, 493–494
common centroid, 103–104, 108, 113
common-drain amplifier, 122–124
common-gate amplifier, 124–127, 138, 179–181, 198, 325–327
common-mode feedback (CMFB), 283285, 288–293, 432
common-mode rejection, 281, 284, 451, 454
common-source amplifier, 36–38, 120–122, 132, 138, 166–170, 172, 174–179, 198, 230, 233–234, 243–244, 265–266, 288, 292, 395
comparator, 413–439, 524, 585, 595–599, 646–647, 651–653, 656, 658, 660–662, 667–678, 681–683, 685–686, 707, 723
hysteresis, 414–415
latched, 426–435, 437–439, 596, 681, 685–686
noise, 413–414
offset, 413–414, 417–418, 430–431
phase, *see* phase detector
preamplifier, 426–427, 430–433, 436–439
specifications, 413–415
compensation, *see* frequency compensation
complex filter, *see* filter, complex
complex signal, 525–527
computer aided design (CAD), 74, 86, 93–94
constant transconductance, 269, 307–310, 315, 327, 519–520
continuous-time filter, *see* filter, continuous-time
conversion time, 617, 650, 652
correlated double-sampling (CDS), 252, 558, 593–594
current mirror:
advanced, 261–267, 309–310
bipolar, 349–350, 360
cascode, *see* cascode, current mirror
CMOS, 101, 118–120, 122, 124, 136–138, 261–267, 275–27, 284–286, 307–310
ideal,101, 118–120, 122, 124, 136–138, 275–277, 284–286, 307–309
source-degenerated, *see* source degeneration, current mirror
symbol, 267
wide-swing, *see* wide-swing current mirror
current-mode D/A, *see* D/A converter, current mode
cyclic A/D, *see* A/D converter, algorithmic
Czochralski method, 73

D

D/A converter:
binary-scaled, 628–634, 640, 642
charge-redistribution, 632–633, 635–636, 642
current-mode, 628, 633–637
decoder-based, 623–628, 642
hybrid, 623, 640–643
ideal, 606–608, 611
Nyquist-rate, 606, 623–642
oversampling, 606, 623, 696, 701–702, 713–714, 723–725, 727–729, 732–734
resistor-string, 623–267, 658
successive approximation, 652
thermometer-code, 623, 634–642, 727–728
decibels (dB), 146–147
dBm, 365
decimation, 702, 711, 714–718, 720
decoder-based D/A, *see* D/A converter, decoder-based
delayed integrator, *see* integrator, delayed
delay-free integrator, *see* integrator, delay-free
delta-sigma modulation, 696, 702–714, 718–732
bandpass, 721–722
error-feedback, 709–710
first-order, 704–707
idle tones, 725–727
interpolative, 703–704, 718–719
multi-bit, 727–730
multi-stage (MASH), 719–720
second-order, 706–708
stability, 709, 718–720, 722–723, 726, 727, 729, 730
third-order, 730–732
depletion capacitance, 4, 6–8, 11–13, 32, 36, 52
bipolar, 336, 341, 346
depletion region:
bipolar, 333–334, 357–358
capacitor, 58, 110
MOS, 16–17, 19–22, 48, 64–65, 67, 166
pn junction, 3–7, 10–11, 13, 63, 395
resistor, 55
depletion transistors, 14, 16, 310
design rule, 87, 90
differential nonlinearity (DNL), 616, 618–621, 628, 634, 636
differential pair:
bipolar, 353–356
layout, 103–104
mismatch, 102
MOS, 102–104, 135–138, 193–197, 303–304
diffusion, 3, 42, 44, 61–63, 75, 76, 343, 346, 356–359
diffusion capacitance, 11–13, 62–63, 335–336, 341, 359
digital error correction, 659–662, 670, 673–674, 678, 679–680, 689, 728–729

diode bridge, 452, 456–458
diode-connected transistor, 118, 136, 266, 272, 289, 308, 431–432, 483
diode, 2–13, 39, 50, 310–311, 319–320, 335, 378–380, 436, 456–458
Dirac delta function, *see* impulse function
direct frequency tuning, *see* tuning, direct frequency tuning
discrete-time filter, *see* filter, discrete-time
discrete-time signal processing, 537, 543, 553
dither, 726
divider:
  admittance, 151–152
  capacitor, 420, 458, 476
  clock, 738–741, 743, 747, 749, 752, 754–755, 777, 779–80
  current, 642
  resistor, 111, 352
  voltage, 321
dominant-pole:
  approximation, 147, 169, 172, 173, 175, 185, 190, 213, 216–218, 220, 227, 234, 246, 248, 275, 327
  compensation, 254–260, 269, 271, 323, 416, 474
donor, 2, 62, 74, 81, 341
doping, 1–2, 4–7, 11, 16, 23–24, 28, 51–52, 61, 64–65, 74, 76–77, 79, 81, 85, 97, 111, 341, 348, 359
downsampling, 543–544, 717
drain-induced barrier lowering (DIBL), 42, 48
drift, 11, 42, 44, 49, 66–67
droop rate, 444–445
dropout voltage, 322, 324
dual-tone multi-frequency (DTMF), 599
dynamic element matching (DEM), 727–728
dynamic range, 331, 363, 377, 383, 397–398, 404, 445, 495, 509, 557, 577, 580, 583
  data converters, 617–618, 689, 696, 702, 709, 713, 719, 720, 722–723, 727, 731
dynamically matched current source, 638–640, 728

E

Early voltage, 334–335, 345–346, 348
effective gate-source voltage, 17, 25–26, 42, 45, 101, 177, 253, 260, 262–263, 266, 272, 277, 289, 303, 309, 419, 426, 448, 484, 492
effective resolution bandwidth, 618
EKV model, 52
emission coefficient, 50
emitter degeneration, 349–350, 463–464, 480–481, 483, 500–502, 518
emitter follower, 349–353, 433, 437–438, 458–459, 506, 677
enhanced output-impedance:
  current mirrors, 263–267, 285
enhancement transistors, 14, 24
epitaxial, 74, 82, 85, 96, 111, 336, 346, 348
error-feedback, 709–710
Euler's relation, 145

F

Fermi potential, 24, 28, 64
field implants, 32–33, 79–81, 85, 89, 97
fill requirements, 94
filter:
  active-RC, 242, 509–512
  biquad, *see* filter, second-order
  complex, 528–531
  continuous-time, 469–531
  discrete-time, 545–552, 572, 760–761
  finite impulse response (FIR), 550, 715, 717, 720
  first-order, 387–390, 470, 572–577
  G_m-C, 471–509
  infinite impulse response (IIR), 550, 715, 717
  MOSFET-C, 512–516
  noise, 371–375, 387–390
  R-MOSFET-C, 516
  second-order, 470–471, 577–585
  switched-capacitor, 242, 572–585
  transconductance-C, *see* filters, G_m-C
  tuning, 516–524
finite impulse response (FIR) filter, *see* filter, finite impulse response (FIR)
first-order filter, *see* filter, first-order
flash A/D, *see* A/D converter, flash
folded-cascode opamp, 252, 268–274, 279, 283–284, 286, 291
folding A/D, *see* A/D converter, folding
Fourier Transform, 145–148, 154, 369, 446, 525, 540, 595, 649, 758
four-transistor integrator, *see* integrator, four-transistor
free-running frequency, 739–740, 747, 754
frequency compensation:
  bias circuit, 308–309
  common-mode feedback, 286, 291, 474
  comparator, 416–417, 424
  current mirror, 277–280
  feedback amplifier, 218, 254–255, 396, 452, 507, 558, 596
  folded-cascode opamp, 268–269, 271, 274
  output impedance enhancement, 267
  source-follower, 191–193
  two-stage opamp, 245–248, 255–261, 509
  voltage regulator, 323–325
frequency response:
  decimation, 716
  feedback amplifier 209, 212–216, 219, 221
  filter, 510–511, 516, 523), 545–548, 572, 574
  integrating A/D, 649
  noise transfer function, 722
  phase locked loop (PLL), 748–756, 777–781
  sample-and-hold (S/H), 552–554
  single-stage amplifier, 144–197, 264–265
  voltage regulator, 326
frequency synthesis, 738, *see also* phase-locked loop (PLL)
fringing capacitance, 31–32, 35, *see also* overlap capacitance

fully-differential circuits:
A/D converter, 662, 672, 674
comparator, 422, 432
filter, 474–477, 484, 491, 506–507, 510, 512, 514, 528–531, 568, 570, 573, 575–578
opamp, 265, 281–292
oscillator, 768–769
sample-and-hold (S/H), 451, 456, 458

G

G_m-C filter, *see* filter, G_m-C
gain boosting, 263–264, 269
gain error, 447, 586
data converter, 615, 618–19, 648
feedback amplifier, 221, 228–229
gate capacitance, 17–18, 35, 165, 380
gate leakage, 49, 381
Gilbert multiplier, *see* translinear multiplier
graded junction, 7–8, 51, 341
Gummel-Poon model, 345–346

H

half-circuit, 194, 355, 430, 514, 517, 771
harmonic distortion, 55–56
total harmonic distortion (THD), *see* total harmonic distortion
heterojunction bipolar transistor (HBT), 334, 347–348
HiCuM model, 346
hold step, *see* sampling pedestal
hot-carrier effects, 48–49, 85
hybrid converter, *see* D/A converter, hybrid
hysteresis, 414–415, 427, 432, 436–437

I

idle tones, 725–727
image-reject mixer, 525
impact ionization, 48, 263
impulse function, 540, 545, 549
impulse response, 144–146, 545, 550, 715, 773
impulse train, 537, 538, 540
indirect frequency tuning, *see* tuning, indirect frequency tuning
infinite impulse response (IIR), *see* filter, infinite impulse response
ingot, 73–74
input isolation, 445
input offset:
comparator, 413–415, 417–418, 424, 427, 432, 434
differential pair, 102, 104
latch, 430–431
opamp, 242, 252–253, 318–319, 449, 453–454, 586, 588–519, 591, 594, 663
integral nonlinearity (INL), 616, 618–619
integrating A/D, *see* A/D converter, integrating
integrator:
active RC, 510–511, 646–647
comparator, 430
complex, 531
delayed, 566, 568
delay-free, 567–568
four-transistor, 515–516
fully differential, 474–475
G_m-C, 472–475, 492, 506–509, 518, 521
ideal, 154, 469–471, 522, 704, 716–717, 729
MOSFET-C, 512–516
noise analysis, 372–373, 392
opamp, 245–249, 268–269
parasitic-insensitive, 565–569
parasitic-sensitive, 562–565, 573
switched-capacitor, 521, 562–569, 593–594, 706–707, 726–727
two-transistor, 512–514
intercept point, 400–403
intermodulation, 400–404, 507, 516, 688
interpolating A/D, *see* A/D converter, interpolating
intrinsic gain, transistor, 36–38, 45, 121, 177, 217, 234, 242, 269, 331, 341–342
intrinsic speed, transistor, *see* unity-gain frequency
inverted channel, 16, 20–21, 59, 64
inverting opamp configuration, 221, 223–224, 231–233, 256, 279–280, 387–390, 510
ion implantation, 75–77, 79, 81–85, 96–98, 346–347

J

jitter:
A/D converter, 688, 704, 723, *see also* aperture jitter
absolute, 756–758, 762, 775
adjacent period, 761–762, 764
aperture, *see* aperture jitter
cycle-to-cycle, *see* jitter, adjacent period
oscillator, 775–776
P-cycle, 761, 764, 775–776
peaking, 749–750, 753–754
period, 760–762, 764
phase-locked loop (PLL), 777
probability density function, 764–765
transfer function, 749–750, 777–781
junction capacitance, 4, 6–13, 24, 32–36, 39, 40, 42, 50–52, 55–56, 58–59, 81, 88, 90–91, 95, 98–99, 165–166, 278, 282, 336, 341, 345–346, 424–425, 458, 559
junction leakage, 49
junction potential, *see* built-in junction potential

K

kickback, 427, 431, 437–439
Kirk effect, 343–344

L

Laplace transform, 144–151, 539, 541, 544, 552, 740
latch-up, 48, 83, 85, 90, 95–96, 111, 768
latency, 617, 665, 673, 677, 690
layout, 52, 73, 74, 86–96, 103–112, 363
lead compensation, 242, 245, 254–260, 269, 271, 274, 277–280, 293, 396, 417, 507
leakage currents, 42, 49, 52, 78, 81, 263, 640
least significant bit (LSB), 606–611, 613–620, 626–627, 633–635, 641–642, 646, 653, 659–660, 669, 673,675–676,678–680,683,697, 701, 732
level shifter, 266
lightly-doped drain, 85
line regulation, 322, 326
linear settling, 216–217, 249–250, 279–280
linearity:
  data converter, 615–619, 634, 656, 673, 681, 701–702, 712–713, 718, 721, 723–725, 727–729
  feedback, 207–209
  filter, 474, 481–484, 486, 488, 493, 495, 498, 500, 507–510, 514–516
  metrics, 363, 397–404
  sample-and-hold, 445
  translinear circuit, 461–463
  voltage-controlled oscillator (VCO), 769
linearized differential pair, 497–498, 505
local oxidation of silicon (LOCOS), 79–81, 97
loop filter, 738–739, 744, 746–747, 749–755, 777, 779–781
loop gain, 95, 454, 596
  common-mode feedback, 289, 291
  current mirror, 263, 265
  feedback amplifier, 205–208, 210–212, 214–215, 217–218, 221–226, 228–235
  opamp, 255–257
  PLL, 748, 768, 773, 777–778
  voltage regulator, 321, 324, 327
low dropout (LDO) regulator, 324–327
low-voltage circuits:
  bandgap, 319–320
  opamp, 287–288

M

magnitude response, 146–148, 156–158, 168, 198
  feedback amplifier, 209, 211, 219
  filter, 527–528, 546
  jitter peaking, 749–750
  noise transfer function, 372, 374, 387, 405–406
  oversampling converter, 722, 730, 733
metastability, 429, 676
metal-metal capacitor, 59–60, 772
Miller capacitance, 32, 68, 104, 177, 185–186, 243, 248, 509
Miller effect, 169–173, 177–178, 181–183, 186, 198, 356
Miller theorem, 169–173, 176
Mismatch, 96, 98–102, 113, 119, 320, 516, 519
  bipolar, 348–349, 360
  D/A converters, 628, 670, 688–690
  latches, 430, 439
  oversampling converters, 710, 728
missing codes, 616
mobility, 18, 43, 51–52, 61, 66, 85, 98–99, 181, 187, 309
mobility degradation, 44–47, 107, 176–177, 491
moderate inversion, 21, 44, 52
modulator:
  amplitude, 594–596
  delta-sigma, *see* delta-sigma modulation
  frequency, 444, 462, 768
monotonicity, 616, 623, 626, 628, 630, 634, 636, 640–641, 658, 660
MOS capacitor, *see* capacitor, MOS
MOSFET-C filter, *see* filter, MOSFET-C
multiplying D/A (MDAC), 607, 623, 665, 667, 670, 672–673

N

native transistor, 65
nichrome resistors, 85
noise:
  1/f, 252, 284, 370–371, 378, 380–382, 390, 394–395, 418, 436, 558, 588, 589, 591, 593, 594, 632, 774
  1/f noise tangent, 376–377
  bandwidth, 373–375, 377, 382, 392
  comparator, *see* comparator, noise
  factor, 387
  figure, 387
  filtered, 371–377
  flicker, *see* noise, 1/f
  input-referred, 284, 380, 384–387, 391–394, 397, 413–414, 426, 570–571
  opamp, 252, 271, 277, 282, 284, 379, 382, 387–390, 392–394, 571, 593
  pink, *see* noise, 1/f
  power, 282, 364–365, 367, 371, 373, 402–404, 610–611, 696–698, 700, 706, 708–709, 713, 723, 726
  quantization, 606, 609–612, 618, 696–711, 713–717, 719–722, 726–727, 729–730
  shot, 378, 380, 390–391
  thermal, 252, 271, 277, 282, 284, 350, 377–381, 384, 386–387, 390, 392, 394–397, 570–571, 593, 670, 700, 711, 758–759, 780, 780
  white, 369–370, 372–375, 377–378, 381–382, 393, 697, 701, 725–726, 728, 756–757, 760–761, 764, 774
noise shaping, *see* delta sigma modulation
noise spectral density, 367–369, 378, 381–383, 386–387, 392, 758–760, 774–775, 777, 780
noise transfer function (N_{TF}), 703–708, 710, 718–722, 726, 730–731
noninverting opamp configuration, 159, 208, 226–232, 255–256

nonlinearity, 10, 208, 283, 462, 481, 495, 576, 615–616, 618, 656, 689, 701, 712, 727–729, 769
nonoverlapping clocks, 557, 559–561
Nyquist:
  frequency, *see* Nyquist-rate
  plot, 210
  rate, 540, 543–544, 553, 606, 623, 689, 696, 699–700, 711, 713, 715, 717–718
  stability criterion, 210

O

offset:
  amplifier, 159, 252–254, 292, 318–320, 348, 390, 449, 447, 450, 454, 456, 557–559, 585–586, 632
  bias differential pair, 499
  binary coding, 606, 612–613, 652
  cancellation, 318, 417–418, 423, 430–431, 588–594, 663–666
  comparator, 413–415, 417–418, 420, 423–424, 426–427, 430–434, 669–673
  error, 445, 615–616, 618–619, 627, 646, 648, 652, 656, 688–689, 720
  frequency, 763, 773, 775
  mismatch, 102, 104, 688–689
  systematic, 252–254
operational transconductance amplifier (OTA), 268, 293, 452, 472, *see also* opamp
optical proximity correction, 75
oscillator, *see* voltage-controlled oscillator (VCO)
overetching, 96–97, 104–105
overlap capacitance, 31–32, 35, 51, 58–59, 68, 77, 83, 86, 165, 419–421, 586–587
oversampling, 384, 455, 611, 696–709
  converters, 537, 594, 606, 618, 647, 696–732
  ratio, 699–700, 706, 708–709, 713, 718, 721, 723, 727
oxide definition (OD) region, 78, 81, 86, 98, 104

P

parasitic insensitive integrator, 565–568
passivation, 84
peak detector, 523, 594, 596
permittivity, 4, 18, 56, 58–59, 66
phase detector, 521, 738–739, 741–746, 747–748, 777
  multiplier, 444, 462, 741–743
  charge pump, 743–746, 752, 755
phase margin:
  common-mode feedback, 291
  current mirror opamp, 277–279
  feedback amplifier, 211–215, 217–221, 228–229, 279, 557–558
  folded-cascode opamp, 270–271, 274
  phase-locked loop (PLL), 751–753
  two-stage opamp, 249, 254–259
phase noise, 756, 758–763
  loop filter, 779–781
  oscillator, 772–776, 778–779
  phase-locked loop (PLL), 777–781
  spur, 762
phase response:
  Bode plot, 157–159, 163
  feedback amplifier, 209–210, 214, 220, 227
  integrator, 507, 511, 522
  linear system, 145–146, 156–157, 546
  phase-locked loop (PLL), 751
phase-locked loop (PLL), 738–781
  filter tuning, 521
  jitter and phase noise, 777–781
  small-signal analysis, 748–756
  translinear multiplier, 444, 462
phosphorus, 2, 76–77
photolithography, 74–76, 113
photoresist, 74–76, 78–79, 81–84, 97
pinch-off, *see* active region, MOSFET operation
pipelined A/D, see A/D converter, pipelined
pole-splitting, 248
polysilicon, 14–15, 31, 47, 54, 59, 64, 81–88, 90–91, 93–98, 107–109, 111, 113, 165, 347, 558, 623
polysilicon resistors, 107–109, 623
positive feedback:
  biasing, 308–309
  comparators, 427–434, 438–439, 677
  feedback amplifiers, 211
  latch-up, 95
  oscillator, 599, 768
power supply rejection, 322, 324–327, 447, 674, 769
preamplifier, 395–396, 426–427, 430–433, 436–439, 618, 675–678
probability density, 610, 764
process-voltage-temperature (PVT), 99
proximity effects, 96–97
PSP model, 52
punch-through, 48–49

Q

Q-factor:
  feedback amplifier, 219
  filter, 519, 522–524, 577–584, 599
  frequency response, 155–156, 164, 189–191
  phase-locked loop (PLL), 749–756
quantization, 608, 614, 678–679
  noise 606, 609–611, 618, 696–704, 706, 708–723, 726–727, 729–731
quantizer overloading, 609, 704

R

RAMDAC, 637
recombination, 358
reconstruction filter, *see* smoothing filter
rectifier circuit, 95, 594, 596–597
reference circuit, 302, 305–306

voltage, *see* bandgap
current, 320–321
residue, 678
resistance reflection rule, 351–352, 355
resistor string bowing, 675, 677
resistor-string D/A, *see* D/A converter, resistor-string
resolution:
-bandwidth, 368–369, 618
comparator, 415, 418, 420, 424, 426–427, 433, 437
data converter, 611, 614–615, 617–618, 669, 673–674, 696
effective, 617–618
frequency, 755
lithographic, 75, 94
resonator, 718–719, 721, 738
return-to-zero (RZ) coding, 724–725
root mean square (rms), 363, 364, 365, 369, 759
root spectral density, 368, 370, 372, 375, 377–378

S

silicide, 82, 85, 108, 109
sampling pedestal (or hold step), 444–445, 448, 451, 464
sample-and-hold (S/H), 49, 384, 444–451, 464, 675, 711
bipolar, 456–459
BiCMOS, 456–460
CMOS, 452–456
sample-and-hold response, 552–554
sampling time uncertainty, *see* aperture jitter
saturation region, *see* active region, MOSFET operation
scale current, 11, 311, 333, 461
Schottky diodes, 3, 12–13, 458
second-order filter, *see* filter, second-order
self-aligned, 83, 347
semiconductor resistor, 57
settling time, 215–217, 221, 249–250, 267, 279–280, 520, 567, 577, 579, 617, 623–624, 659, 751, 753, 769
shallow-trench isolation (STI), 78, 80, 85, 98
sheet resistance, 54–56, 85, 108–109
short-channel effects, 22–23, 48–49, 51–52, 57, 131, 182, 261, 263, 267, 696
sidewall capacitance, 32–33, 35, 50, 52, 88
sign magnitude, 606, 612–613
signal transfer function, 703–706, 709
signal-flow-graph (SFG), 525–528, 531, 569–570, 572, 577–580, 582, 584, 588
signal-to-noise and distortion ratio (SNDR)
signal-to-noise ratio (SNR), 282, 363, 365, 384–385, 387, 389–390, 394, 402, 404, 516, 606, 611, 700–701, 706, 709, 727
signal-to-quantization noise ratio (SQNR), 611, 618, 700–701, 706, 708, 709, 718
silicon dioxide, 15, 56, 59, 74, 79
silicon-controlled rectifier (SCR), 95
sinc:
filter, 715–717
response, 553, 554
slew rate, 245, 249–250, 281, 444–445, 453, 557–558, 590
current mirror opamp, 276–278, 285–286
folded-cascode opamp, 269, 272, 274, 282
two-stage opamp, 250–252
small-signal:
analysis, 117–138, 167, 172–197, 204–205, 220, 222–227, 233–234, 247–248, 270–271, 322–326, 351–356
bipolar transistor model, 338–345
capacitance, 6, 63
capacitor model, 58–60
diode model, 12–13, 39
MOSFET model, 25–37, 40–42, 45, 47, 165–166
MOSFET-C filter, 513–516
oscillator, 767, 771
phase-locked loop (PLL) model, 748–756, 777
resistor model, 54–55
transconductor, 484–488, 493
smoothing filter, 557, 696, 702, 713, 721
source degeneration:
current mirror, 127–129, 180, 308
differential pair, 289, 480–491
source-follower, 65, 122–124, 187–193, 243–244, 252, 290, 677
spot noise, 387
spurious-free dynamic range (SFDR), 402–404
square-law, 1, 21, 26, 42, 45–47, 246, 262, 273, 303, 306, 493–495, 499
stability:
bias circuit, 308
common-mode feedback, 474
delta-sigma modulator, 709, 718–720, 722–723, 726, 727, 729, 730
discrete-time filters, 548–550
feedback amplifier, 208–212, 221
first-order system, 214–215
opamp, 187, 197, 243, 249, 255, 263, 269, 270, 417, 449
phase-locked loop (PLL), 749, 752, 754
second-order system, 217
third- and higher-order systems, 220
transimpedance amplifier, 396
voltage regulator, 324, 327
static nonlinearity, 616, 728
step response:
common-mode feedback, 291
feedback amplifier, 212, 219, 248, 249
filter, 519, 522, 524
first-order, 149–153
phase-locked loop (PLL), 749–750
second-order, 156–157, 163–164
source-follower, 189
strip resistor, 54–57
strong inversion, 20–21, 29, 42–44, 46–47, 52
substrate noise, 85, 106, 109, 111, 282, 676
subthreshold:
leakage, 17, 43, 49
operation, 21, 42–44, 47, 101, 121, 179, 246, 273, 276–277, 394
slope, 43–44

successive approximation A/D, *see* A/D converter, successive approximation
successive approximation register (SAR), 652
summer, 469–470, 472–473, 510
supply noise, 306, 325, 327, 447, 521–522, 676
switched capacitor:
  charge-redistribution , 632–633, 635–636, 642, 653–659
  circuit, 107, 111–112, 216, 242, 249, 272, 281, 318–319, 520–521, 557–600, 726–727
  common-mode feedback, 291–292
  comparator, 417, 422, 434
  delta-sigma modulator, 706–707, 724, 726–727, 732
  filter, 469, 537, 571–585
  gain, 588–593, 627, 629, 632
  modulator, 594–595
  multiplying D/A, 672
  noise, 384, 570–571
  oscillator, 596–600
  peak detector, 596
  rectifier, 595–596
  sample-and-hold (S/H), 446–456
switched emitter follower, 458–459

T

T model, 130, 135–136, 193–194, 340, 351, 355, 481, 488
tanh relationship, 354, 462
Taylor series, 22, 283, 398, 576
thermometer-code:
  D/A converter, *see* D/A converter, thermometer-code
  flash converter, 673, 676
  folding A/D converter, 683
threshold voltage, 16, 20, 24, 43–44, 48, 51–52, 60, 64–65, 78–79, 81, 85, 97–100, 132, 263, 304, 307, 310, 436, 448, 490, 493–494, 499, 559, 585–586, 653, 656, 674, 680
total harmonic distortion (THD), 398–400, 404, 502, 505, 516
track-and-hold, *see* sample-and-hold (S/H)
transconductance-C filter, *see* filter, G_m-C
transconductor:
  BiCMOS, 506–509
  bipolar, 480–483, 500–506
  differential, 474–475
  ideal, 471–475
  MOSFET, 479–480, 482, 484–500
  tuning, 517–520
transimpedance amplifiers, 231–234, 395–397
transit time, 11–12, 50, 64, 336, 343–346, 348, 359, 425
translinear gain, 444, 460–462
translinear multiplier, 444, 462–464, 507–509
triode:
  MOS resistor, 57–58, 193, 257, 260–261, 483, 522, 769–770
  MOSFET-C circuits, 512–515
  operation, 20–21, 23, 33–35, 40, 66–68, 381, 383
  switch, 418–420, 585
  transconductors, 484–492, 506–507
tuning, *see* filter, tuning
twin-well process, 78, 84
two-stage opamp, 136, 138, 196–197, 242–252, 255–261, 283, 288, 291–293, 386, 393–394
two-transistor integrator, *see* integrator, two-transistor

U

ultraviolet light, 74–75
unity-gain frequency:
  amplifier, 154, 264–265
  BJT, 342–343, 348, 360
  feedback amplifier, 210–220, 227
  integrator, 472–473, 510–511, 518
  inverter, 429, 557–558, 768–769
  multi-stage comparators, 424–425
  opamp, 245–247, 251–252, 254–260, 264–265, 270, 272–273, 276–280, 292–293, 388–389, 416, 452
  transfer function, 154
  transistor, 37–38, 187
upsampling, 543–544, 554

V

varactor, 58, 771–772
velocity saturation, 44–45, 491, 495
voltage-controlled oscillator (VCO), 521, 596–599, 738–740, 765–766, 770–776, 778–781
  crystal, 765–766
  LC, 765–766, 771–773
  RC, 765–766
  relaxation, 596–599, 765–766
  ring, 674, 765–770, 772–773
  switched-capacitor, 765–766
voltage regulator, 302, 306–307, 321
Volterra series, 398

W

weak inversion, 42–44, 319, 394
well, 6, 14–15, 24, 59
wide-swing cascade, 268–270, 285, 309–310
wide-swing current mirror, 261–263, 309
  enhanced output impedance, 266–267, 285
Wiener–Khinchin theorem, 369

X

X-rays, 75

Z

zero-value time constant, 173–176, 179–183, 190, 623–624, 659

Constants

$q = 1.602 \times 10^{-19} C$	$k = 1.38 \times 10^{-23} JK^{-1}$
$n_i = 1.1 \times 10^{16}$ carriers/m^3 at $T = 300°K$	$\varepsilon_0 = 8.854 \times 10^{-12}$ F/m
$K_{ox} \cong 3.9$	$K_s \cong 11.8$

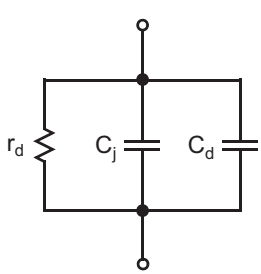
Reverse-Biased Diode (Abrupt Junction)

$C_j = \dfrac{C_{j0}}{\sqrt{1 + \dfrac{V_R}{\Phi_0}}}$	$Q = 2C_{j0}\Phi_0\sqrt{1 + \dfrac{V_R}{\Phi_0}}$
$C_{j0} = \sqrt{\dfrac{qK_s\varepsilon_0}{2\Phi_0}\dfrac{N_D N_A}{N_A + N_D}}$	$C = \sqrt{\dfrac{qK_s\varepsilon_0 N_D}{2\Phi_0}}$ if $N_A >> N_D$
$\Phi_0 = \dfrac{kT}{q}\ln\left(\dfrac{N_A N_D}{n_i^2}\right)$	

Forward-Biased Diode

$I_D = I_s e^{V_D/V_T}$	$I_s = A_D q n_i^2\left(\dfrac{D_n}{L_n N_A} + \dfrac{D_p}{L_p N_D}\right)$
$V_T = \dfrac{kT}{q} \cong 26$ mV at $300°K$	

Small-Signal Model of Forward-Biased Diode



$r_d = \dfrac{V_T}{I_D}$	$C_T = C_d + C_j$
$C_d = \tau_T\dfrac{I_D}{V_T}$	$C_j \cong 2C_{j0}$
$\tau_T = \dfrac{L_n^2}{D_n}$	

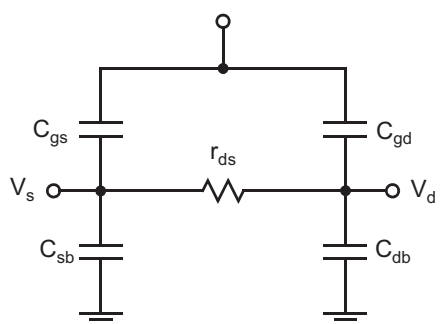
MOS Transistor Triode Region ($V_{GS}>V_{tn}$, $V_{DS}\leq V_{eff}$)

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[V_{DS}(V_{GS}-V_{tn})-\frac{V_{DS}^2}{2}\right]$$

$V_{eff} = V_{GS}-V_{tn}$	$V_{tn} = V_{tn-0}+\gamma(\sqrt{V_{SB}+2\phi_F}-\sqrt{2\phi_F})$
$\phi_F = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right)$	$\gamma = \frac{\sqrt{2qK_{si}\varepsilon_0 N_A}}{C_{ox}}$

$$C_{ox} = \frac{K_{ox}\varepsilon_0}{t_{ox}}$$

Small-Signal Model in Triode Region (for $V_{DS} \ll V_{eff}$)



$$r_{ds} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}}$$

$$C_{gd} = C_{gs} \cong \frac{1}{2}WLC_{ox} + WL_{ov}C_{ox}$$

$$C_{sb} = C_{db} = \frac{C_{j0}(A + WL/2)}{\sqrt{1+\frac{V_{sb}}{\Phi_0}}}$$

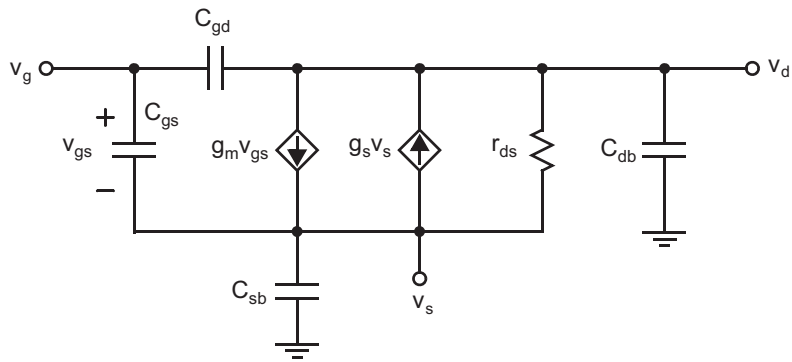
MOS Transistor Active (or Pinch-Off) Region ($V_{GS}>V_{tn}$, $V_{DS}\geq V_{eff}$)

$$I_D = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS}-V_{tn})^2[1+\lambda(V_{DS}-V_{eff})]$$

$\lambda \propto \frac{1}{L\sqrt{V_{DS}-V_{eff}+\Phi_0}}$	$V_{tn} = V_{tn-0}+\gamma(\sqrt{V_{SB}+2\phi_F}-\sqrt{2\phi_F})$

$$V_{eff} = V_{GS}-V_{tn} = \sqrt{\frac{2I_D}{\mu_n C_{ox}W/L}}$$

Small-Signal Model in Active Region



$g_m = \mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}$	$g = \frac{2I_D}{V_{eff}}$	$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$	$g_s = \frac{\gamma g_m}{2\sqrt{V_{SB} + \lvert 2\phi_F \rvert}}$
$r_{ds} = \frac{1}{\lambda I_{D-sat}} \cong \frac{1}{\lambda I_D}$	$g_s \cong 0.2 g_m$	$\lambda = \frac{k_{rds}}{2L\sqrt{V_{DS} - V_{eff} + \Phi_0}}$	$k_{rds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}}$
$C_{gs} = \frac{2}{3}WLC_{ox} + WL_{ov}C_{ox}$	$C_{gd} = WL_{ov}C_{ox}$	$C_{sb} = (A_s + WL)C_{js} + P_s C_{j-sw}$	$C_{js} = \frac{C_{j0}}{\sqrt{1 + V_{SB}/\Phi_0}}$
$C_{db} = A_d C_{jd} + P_d C_{j-sw}$	$C_{jd} = \frac{C_{j0}}{1 + V_D/\Phi_0}$		

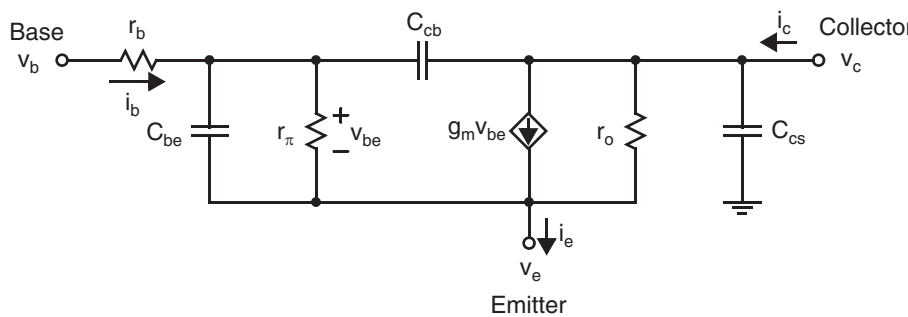
MOSFET parameters representative of various CMOS technologies

Technology	**0.8 μm**		**0.35 μm**		**0.18 μm**		**45 nm**	
	NMOS	**PMOS**	**NMOS**	**PMOS**	**NMOS**	**PMOS**	**NMOS**	**PMOS**
μC_{ox} (μ A/V^2)	90	30	190	55	270	70	300	90
V_{t0} (V)	0.80	–0.90	0.57	–0.71	0.45	–0.45	0.45	–0.45
$\lambda \cdot L$ (μm/V)	0.12	0.08	0.16	0.16	0.08	0.08	0.10	0.15
C_{ox} (fF/μm^2)	1.9	1.9	4.5	4.5	8.5	8.5	25	25
t_{ox} (nm)	20	20	8	8	5	5	1.2	1.2
n	1.5	1.5	1.8	1.7	1.6	1.7	1.85	1.85
θ (1/V)	0.6	0.6	1.5	1.0	1.7	1.0	2.3	2.0
m	1.0	1.0	1.8	1.8	1.6	2.4	3.0	3.0
$C_{ov}/W = L_{ov}C_{ox}$(fF/μm)	0.20	0.20	0.20	0.20	0.35	0.35	0.50	0.50
$C_{db}/W \approx C_{sb}/W$(fF/$\mu$m)	0.50	0.80	0.75	1.10	0.50	0.55	0.45	0.50

Bipolar Active Transistor

$I_C = I_{CS} e^{V_{BE}/V_T}$	$V_T = \frac{kT}{q} \cong 26 \text{ mV at } 300\ ^\circ\text{K}$
For more accuracy, $I_C = I_{CS} e^{V_{BE}/V_T}\left(1 + \frac{V_{CE}}{V_A}\right)$	
$I_{CS} = \frac{A_E q D_n n_i^2}{W N_A}$	$I_B = \frac{I_C}{\beta}$
$I_E = \left(1 + \frac{1}{\beta}\right) I_C = \frac{1}{\alpha} I_C = (\beta + 1) I_B$	$\beta = \frac{I_C}{I_B} = \frac{D_n N_D L_p}{D_p N_A W} \cong 2.5 \frac{N_D L_p}{N_A W}$
$\alpha = \frac{\beta}{1 + \beta}$	

Small-Signal Model of an Active BJT



$g_m = \frac{I_C}{V_T}$	$r_\pi = \frac{V_T}{I_B} = \frac{\beta}{g_m}$	$r_e = \frac{\alpha}{g_m}$	$r_o = \frac{V_A}{I_C}$	$g_m r_o = \frac{V_A}{V_T}$
$C_{be} = C_j + C_d$	$C_j \cong 2A_E C_{je0}$	$C_d = \tau_b \frac{I_C}{V_T} = g_m \tau_b$	$C_{cb} = \frac{A_C C_{jc0}}{\left(1 + \frac{V_{CB}}{\Phi_{c0}}\right)^{1/3}}$	$C_{cs} = \frac{A_T C_{js0}}{\left(1 + \frac{V_{CS}}{\Phi_{s0}}\right)^{1/2}}$